Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado

Facultad de Ciencias Matemáticas

Unidad de Posgrado

Optimización del clasificador "Naive Bayes" usando árbol de decisión C4.5

TESIS

Para optar el Grado Académico de Magíster en Estadística

AUTOR

Carlos ALARCÓN JAIMES

ASESOR

Mg. Emma Norma CAMBILLO MOYANO

Lima, Perú

2015

Referencia bibliográfica

Alarcón, C. (2015). *Optimización del clasificador "Naive Bayes" usando árbol de decisión C4.5*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

ACTA DE SUSTENTACIÓN DE TESIS DE GRADO ACADÉMICO DE MAGÍSTER

Siendo las, ... horas del día viernes 16 de enero de 2015, en la Sala de Profesores de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos, el Jurado Evaluador de Tesis, Presidido por el Mg. Wilfredo Domínguez Cirilo e integrado por los siguientes miembros, Dr. Erwin Kraenau Espinal (Jurado Evaluador), Mg. Rosario Bullón Cuadrado (Jurado Evaluador), Mg. Olga Lidia Solano Dávila (Jurado Informante) y la Mg. Emma Norma Cambillo Moyano como Miembro Asesor, se reunieron para la sustentación de la tesis titulada: "OPTIMIZACIÓN DEL CLASIFICADOR "NAIVE BAYES" USANDO ARBOL DE DECISIÓN C4.5" presentada por el Bachiller Carlos Alarcón Jaimes, para optar el Grado Académico de Magíster en Estadística.

Luego de la exposición del graduando, los Miembros del Jurado hicieron las preguntas correspondientes, así como las observaciones e inquietudes acerca del trabajo de tesis, a las cuales el Bachiller Carlos Alarcón Jaimes respondió con acierto y solvencia, demostrando pleno conocimiento del tema.

A continuación se realizó la calificación correspondiente, según tabla adjunta, resultando el Bachiller Carlos Alarcón Jaimes aprobado con el calificativo de ...17..... ....DIECISIETE (MUY BUENO)

Habiendo sido aprobada la sustentación de la Tesis, el Jurado Evaluador recomienda para que el Consejo de Facultad apruebe el otorgamiento del grado académico de **Magíster en Estadística** al Bachiller Carlos Alarcón Jaimes.

Siendo las 10:55 horas, se levantó la sesión, firmando para constancia la presente Acta.

Dr. Erwin Kraenau Espinal
Miembro

Mg. Wilfredo Domínguez Cirilo
Presidente

Mg. Rosario Bullón Cuadrado
Miembro

Mg. Olga Lidia Solano Dávila
Miembro

Mg. Emma Norma Cambillo Moyano
Miembro Asesor

Para mis
padres y familia.

AGRADECIMIENTOS

A mis profesores del área de Estadística del Posgrado de la Facultad de Ciencias Matemáticas de la UNMSM por todas las enseñanzas brindadas.

A los profesores que contribuyeron con sus sugerencias y aportes en la realización de la presente tesis.

Asimismo, un agradecimiento especial a la profesora Mg. Emma N. Cambillo Moyano como Asesora de la tesis, por la contribución al logro de este trabajo y las enseñanzas recibidas.

A todas las personas que participaron en la consecución de esta tesis.

Carlos Alarcón Jaimes

ÍNDICE GENERAL

Lista de Cuadros

Lista de Figuras

RESUMEN

El clasificador Naive Bayes es uno de los modelos de clasificación más efectivos, debido a su simplicidad, resistencia al ruido, poco tiempo de procesamiento y alto poder predictivo. El clasificador Naive Bayes asume una fuerte suposición de independencia entre las variables predictoras dada la clase, lo que generalmente no se cumple. Muchas investigaciones buscan mejorar el poder predictivo del clasificador relajando esta suposición de independencia, como el escoger un subconjunto de variables que sean independientes o aproximadamente independientes.

En este trabajo, se presenta un método que busca optimizar el clasificador Naive Bayes usando el árbol de decisión C4.5. Este método, selecciona un subconjunto de variables del conjunto de datos usando el árbol de decisión C4.5 inducido y luego aplica el clasificador Naive Bayes a estas variables seleccionadas. Con el uso previo del árbol de decisión C4.5 se consigue remover las variables redundantes y/o irrelevantes del conjunto de datos y escoger las que son más informativas en tareas de clasificación, y de esta forma mejorar el poder predictivo del clasificador. Este método es ilustrado utilizando tres conjuntos de datos provenientes del repositorio UCI , *Irvin Repository of Machine Learning databases de la Universidad de California* y un conjunto de datos proveniente de la Encuesta Nacional de Hogares del Instituto Nacional de Estadística e Informática del Perú, ENAHO – INEI, e implementado con el programa WEKA.

Palabras claves: Redes bayesianas, clasificador bayesiano, Naive Bayes, árbol de decisión C4.5 .

ABSTRACT

The Naive Bayes classifier is one of the most effective classification models, due to their simplicity, resistance to noise, little processing time and high predictive power. The Naive Bayes classifier assumes a strong assumption of independence between the predictor variables, which generally is not met. Many studies seek to improve the predictive power of classifier relaxing this assumption of independence, as the selection of a subset of variables that are independent or approximately independent.

In this paper, a method that seeks to optimize the Naive Bayes classifier using the C4.5 decision tree is presented. This method selects a subset of variables in the data set using the C4.5 decision tree induced for go to apply Naive Bayes classifier to these selected variables. With the previous use of the decision tree C4.5 is achieved removing redundant and / or irrelevant variables in the dataset and choose those that are more informative in classification tasks, and thus improve the predictive power of the classifier. This method is illustrated using three data sets from the UCI repository, *Irvin Repository of Machine Learning databases at the University of California* and a set of data from the National Household Survey of the National Institute of Statistics and Informatics of Peru, ENAHO - INEI , and implemented in WEKA program.

Keywords: Bayesian networks, Bayesian classifier, Naive Bayes, C4.5 decision tree.

INTRODUCCIÓN

Los métodos de clasificación son aplicados en diversas áreas, por ejemplo, en la concesión de créditos, diagnóstico médico, campañas de marketing, clasificación de textos, etc. Los métodos más conocidos que se han propuesto están basados en el análisis discriminante, regresión logística, árboles de decisión, redes bayesianas y redes neuronales.

Un clasificador es una función que asigna (clasifica) una observación en una de las clases predefinidas.

Los clasificadores bayesianos están basados en las redes bayesianas [31], [13]. Éstas son modelos gráficos probabilísticos que permiten modelar de una forma simple y precisa la distribución de probabilidad subyacente a un conjunto de datos. Esto es, las redes bayesianas son representaciones gráficas de las relaciones de dependencia e independencia entre las variables presentes en el conjunto de datos que facilitan la comprensión e interpretabilidad del modelo.

Entre los modelos de clasificación, los clasificadores bayesianos vienen teniendo buenos resultados y son tan competitivos como los árboles de decisión y redes neuronales [37], y se están usando exitosamente en muchas aplicaciones relacionadas con la clasificación [2], [11], [1].

Principalmente el clasificador Naive Bayes (llamado también, clasificador bayesiano simple) es uno de los más efectivos modelos de clasificación [23], [26], [8], debido a su simplicidad, resistencia al ruido, poco tiempo para el procesamiento y alto poder predictivo (tasa de acierto).

El clasificador Naive Bayes asume una fuerte suposición de independencia entre las variables predictoras dada la clase, lo que generalmente no se cumple, ya que en datos

del mundo real no siempre se puede satisfacer la suposición de independencia entre las variables.

El rendimiento del clasificador puede decrecer en presencia de variables correlacionadas, por lo que la implementación de un método de selección de variables, que remueva las variables redundantes y/o irrelevantes, usando el árbol de decisión C4.5 generará una mejora en el poder predictivo del clasificador y una simplificación del modelo.

En este trabajo, se presenta un método que busca optimizar el clasificador Naive Bayes usando el árbol de decisión C4.5. Este método, selecciona un subconjunto de variables que son obtenidos a partir del árbol de decisión C4.5 inducido del conjunto de datos y luego aplica el clasificador Naive Bayes a estas variables obtenidas. Con el uso previo del árbol de decisión C4.5 se consigue remover las variables redundantes y/o irrelevantes (eliminar las variables altamente correlacionadas) del conjunto de datos, y escoger las que son más informativas en tareas de clasificación, y de esta forma mejorar el poder predictivo del clasificador.

Muchas investigaciones se han desarrollado buscando relajar la suposición de independencia del clasificador Naive Bayes [30], [36], [24], con el fin de mejorar el poder predictivo del clasificador.

En general, estas investigaciones principalmente se pueden dividir en dos grupos. Unos intentan escoger un subconjunto de variables que sean independientes (o aproximadamente independientes), para luego aplicar el clasificador solo con estas variables [27], [34]. El otro grupo busca utilizar las relaciones que se puedan presentar entre las variables, modificando el clasificador, lo que se convertiría en un clasificador Naive Bayes Aumentado que incluya esas relaciones (Semi-Naive Bayes SNB [24], búsqueda de independencias [29], el Naive Bayes Aumentado en Arbol TAN [13], SuperParent TAN SP-TAN [21], Lazy Bayes Rule LBR [40]).

El presente trabajo ha sido estructurado como sigue. En el capítulo I, se describen los fundamentos de las redes bayesianas. En el capítulo II, se presenta las características de un clasificador bayesiano en general, especificando al clasificador Naive Bayes. En el capítulo III, se presenta los procedimientos para la inducción de los árboles de decisión.

El capítulo IV, concierne a la optimización del clasificador Naive Bayes, este procedimiento es ilustrado utilizando tres conjuntos de datos provenientes del repositorio UCI , *Irvin Repository of Machine Learning databases de la Universidad de California* y un conjunto de datos proveniente de la Encuesta Nacional de Hogares del Instituto Nacional de Estadística e Informática del Perú, ENAHO – INEI, e implementado con el programa WEKA.

Con el propósito de mejorar el rendimiento del clasificador Naive Bayes, mostrando un método alternativo de selección de variables, y además de contribuir al conocimiento de la metodología bayesiana aplicada al área de clasificación, que es poco tratada en nuestro medio, se define el presente estudio: OPTIMIZACIÓN DEL CLASIFICADOR "NAIVE BAYES" USANDO ÁRBOL DE DECISIÓN C4.5 .

CAPÍTULO I

REDES BAYESIANAS

Introducción

Las redes bayesianas se desarrollaron a finales de los años 80 a partir de una serie de trabajos entre los que destacan los de [31], [25].

La red bayesiana es un grafo acíclico dirigido (estructura gráfica) que representa las relaciones probabilísticas de un conjunto de variables y que permite realizar inferencia probabilística de éstas variables.

Una red bayesiana consiste de dos partes, la construcción de la red Bayesiana y la inferencia en la red.

En la construcción de la red bayesiana, hay dos cosas por considerar, la especificación de la estructura gráfica de la red y la especificación de los parámetros de las distribuciones de probabilidad que se definen en la red.

Habiéndose construido la red bayesiana para un conjunto de variables aleatorias, ésta representa nuestro "modelo" para las variables. Cuando se tiene información disponible de algunas de las variables, podemos usar este "modelo" para realizar inferencia acerca de las variables no observables en la red. La inferencia en redes bayesianas es realizado usando el Teorema de Bayes.

1.1 Definición de la red bayesiana

Una red bayesiana para un conjunto de variables aleatorias $\boldsymbol{X = (X_1, X_2, ..., X_n)}$ es un par $\boldsymbol{B = (G, P)}$. Donde $\boldsymbol{G}$, es un grafo acíclico dirigido cuyos nodos corresponden a las variables aleatorias $\boldsymbol{X_1, X_2, ..., X_n}$, y cuyos arcos representan las relaciones de dependencia directa entre las variables. Y $\boldsymbol{P}$, representa el conjunto de distribuciones de probabilidad para cada variable en la red $\boldsymbol{B}$ [13], [31].

En la red $\boldsymbol{B}$ se asume suposiciones de independencia: esto es, cada variable $\boldsymbol{X_i}$ es independiente de sus no descendentes dado sus padres. Estas suposiciones en la red es la llamada *Condición de Markov* en la red bayesiana, y se emplea en la factorización de la distribución de probabilidad conjunta de las variables del conjunto $\boldsymbol{X}$.

Una red bayesiana $\boldsymbol{B}$ define una única distribución de probabilidad conjunta sobre $\boldsymbol{X}$ dado por

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i / pa(X_i)) \tag{1.1}$$

donde $pa(X_i)$ representa los padres de $\boldsymbol{X_i}$ en la red $\boldsymbol{B}$.

$P(x_i / pa(X_i))$ es la probabilidad condicional de $\boldsymbol{X_i}$ dado $pa(X_i)$

Un ejemplo de red bayesiana para cinco variables $\boldsymbol{(X_1, X_2, ..., X_5)}$ se muestra en la Figura 1.1, los nodos del grafo $\boldsymbol{G}$ (estructura gráfica) representan las variables y los arcos las relaciones de dependencia entre las variables, sobre cada variable se especifica una distribución de probabilidad marginal o condicional, y el conjunto de todas las distribuciones de probabilidad se representa por $\boldsymbol{P}$.

En la red, la variable que es apuntada por el arco es dependiente de la variable que está en el origen de ese arco. El arco entre X_2 y X_4 indica que la variable X_4 es dependiente de la variable X_2, también se dice, que X_2 es padre de X_4 , o que X_4 es

hijo de X_2 . A partir de la relación existente entre estas variables, en el nodo de la variable X_4 , se especifica la distribución de probabilidad condicional $P(x_4 / x_2)$.

También para las variables X_1 y X_5, X_5 es llamado descendente de X_1 si hay un camino de X_1 a X_5 .

La estructura de la red nos da información sobre las dependencias probabilísticas entre las variables.

La red también representa las independencias condicionales de una variable (o conjunto de variables) dada otra variable(s).

En la red bayesiana $\{X_4\}$ *es condicionalmente independiente de* $\{X_1, X_3, X_5\}$ *dado* $\{X_2\}$ (Por la suposición que se realiza en la red, esto es, por la Condición de Markov: cada variable $\boldsymbol{X_i}$ es independiente de sus no descendentes dados sus padres)

con lo cual: $P(x_4 / x_1, x_3, x_5, x_2) = P(x_4 / x_2)$.

(Se sabe para dos variables, X e Y son independientes si $P(x / y) = P(x)$,

también, X e Y son independientes dado Z si $P(x / y, z) = P(x / z)$)

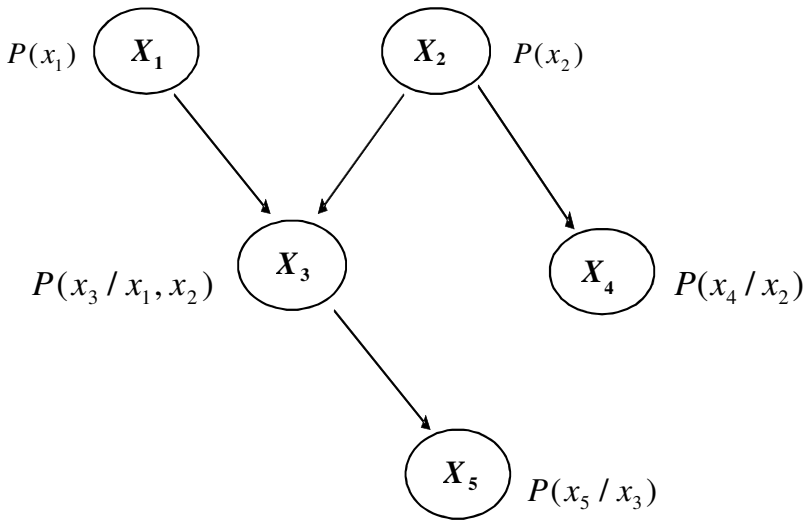


Figura 1.1.- Ejemplo de red bayesiana para cinco variables

La distribución de probabilidad conjunta en esta red es: (empleando la ecuación 1.1)

$$P(x_1, x_2, ..., x_5) = P(x_1)P(x_2)P(x_3 / x_1, x_2)P(x_4 / x_2)P(x_5 / x_3) \qquad (1.2)$$

1.2 Construcción de la red bayesiana

Cuando se construye una red bayesiana, hay dos cosas por considerar, la especificación de la estructura de la red y la especificación de los parámetros de las distribuciones de probabilidad que se definen en la red.

La construcción de la red bayesiana puede ser realizado: a partir del conocimiento de expertos, a partir de datos (conjunto de datos), o puede ser obtenido usando ambas técnicas [16].
En lo siguiente, se tratará la construcción de la red bayesiana a partir de un conjunto de datos.

1.2.1 Especificación de la estructura de la red bayesiana

La especificación de la estructura de la red bayesiana a partir de datos, llamada *estimación estructural*, consiste en encontrar la estructura gráfica de red bayesiana que mejor se ajuste a los datos, esto es, la estructura grafica que mejor represente el conjunto de dependencias/independencias presentes en los datos.

En general, en la literatura aparecen dos metodologías para afrontar el problema de estimación estructural:

- Métodos basados en búsqueda y score (puntuación)
- Métodos basados en detección de independencias (pruebas de independencia)

Métodos basados en búsqueda y score

Este método consiste en generar distintos grafos mediante un algoritmo de búsqueda y aplicar a cada uno de ellos una función de medida (puntuación o score) de calidad para decidir que grafo conservar.

La medida es para evaluar que tan "buena" es cada estructura respecto a los datos. El método de búsqueda es el que genera diferentes estructuras hasta encontrar la "óptima", de acuerdo a la medida seleccionada, esto es, encontrar la estructura gráfica que maximice esta métrica.

Existen muchos algoritmos que siguen esta técnica, definidos a partir de la combinación de los dos elementos:
Algoritmo de búsqueda
Medida global de ajuste (puntuación)

Algoritmo de búsqueda del modelo

Intentar una búsqueda exhaustiva por todo el espacio de grafos es sencillamente intratable. Por tanto, es habitual emplear algún mecanismo de búsqueda *heurística* para guiar el problema y así encontrar cada vez mejores redes bayesianas.

Se han planteado en la literatura específica del tema algoritmos de búsqueda para la resolución aproximada de la estimación estructural. Se mencionan tres de los algoritmos más utilizados para la estimación: K2 , B, y un algoritmo basado en ascenso de colinas HB (*hill climbing*) [28].

Algoritmo K2

El algoritmo K2 es propuesto por Cooper y Herskovits (1992), [6]. Se puede considerar el primer algoritmo basado en búsqueda u optimización de una métrica bayesiana y ha sido fuente de inspiración para posteriores trabajos sobre el tema. Este algoritmo utiliza un esquema voraz (sigue un método que consiste en elegir la opción óptima en cada paso local) en su búsqueda de soluciones candidatas cada vez mejores y parte de que las variables de entrada están ordenadas, de forma que los posibles padres de una variable aparecen en el orden antes que ella misma. Esta restricción es bastante fuerte pero fue un estándar en el origen de los primeros trabajos sobre aprendizaje de redes bayesianas. El proporcionarle al algoritmo un orden entre las variables hace que éste tan "sólo" tenga que buscar el mejor conjunto de padres posible de entre las variables predecesoras en el orden.

Algoritmo B

Este algoritmo es también de los primeros en aparecer en la literatura especializada en el tema, y también está basado en la optimización de una métrica de calidad de redes bayesianas. Al igual que el algoritmo K2, éste se basa en un sistema voraz para la construcción de una solución aproximada a partir de la red vacía de enlaces (cada nodo posee inicialmente un conjunto vacío de padres). A diferencia del algoritmo previo, este algoritmo no impone la restricción de proporcionarle como entrada un orden específico entre las variables.

Este algoritmo trata de introducir el arco que mayor ganancia representa con respecto a la red anterior y a la métrica utilizada. El algoritmo se detiene cuando la inclusión de un arco no presenta ninguna ganancia.

Algoritmo HC

Es un algoritmo local de ascenso de colinas (*hill climbing*) por el máximo gradiente basado en la definición de una vecindad. El algoritmo parte de una solución inicial, como podría ser la red vacía de enlaces, u otra cualquiera (por ejemplo, en tareas de clasificación se podría inicializar con una estructura de Naive Bayes). A partir de esta solución se calcula el nuevo valor de la métrica utilizada de todas las soluciones (grafos) vecinas a la solución actual y nos quedamos con el vecino que mejor valor de la métrica resulte. Estos algoritmos, al igual que los anteriores, se aprovechan de la descomponibilidad de las métricas para recalcular sólo las modificaciones que se realizan en los grafos vecinos definidos.

La vecindad clásica que se maneja en este algoritmo es la siguiente: dado un grafo G (solución actual) se denominan grafos vecinos G' a aquellos resultantes de incluir un solo arco a G o borrar un solo arco presente en G o invertir la dirección de un arco presente en G, todo ello sin incluir ciclos dirigidos en G'. El algoritmo parará cuando no exista ningún vecino que pueda mejorar la solución actual (óptimo local).

Medida de puntuación de calidad

Una medida de calidad permite determinar el grado de ajuste de la estructura de la red y los datos.

Una medida de calidad es utilizada para definir dentro de un conjunto de estructuras aquella que mejor se ajusta al conjunto de datos.

Hay varias posibles medidas de ajuste, las dos más comunes son la medida bayesiana y la medida basada en el principio de longitud de descripción mínima (MDL).

Medida bayesiana

La medida de calidad bayesiana se basa en el teorema de bayes. La medida bayesiana asigna a través del cálculo de la probabilidad a posteriori una medida de calidad para cada estructura de red generada por el algoritmo de búsqueda.
Con lo que busca maximizar la probabilidad de la estructura de la red dado los datos $P(G/D)$, que es calculada a partir de la probabilidad a posteriori:

$$P(G/D) = \frac{P(G,D)}{P(D)} = \frac{P(G)P(D/G)}{P(D)} \tag{1.3}$$

Como los datos son siempre los mismos para las distintas redes de un mismo problema, el denominador $P(D)$ de la anterior expresión puede ignorarse. El término $P(G)$ representa la distribución a priori de cada estructura candidata; en muchos casos se utiliza una distribución uniforme por lo que también puede ignorarse, Finalmente, el término $P(D/G)$, es la verosimilitud muestral.

En el caso de redes bayesianas de variables discretas multinomiales, y asumiendo ciertas condiciones acerca de que los parámetros de la estructura de la red tienen una distribución a priori Dirichlet, se puede obtener la siguiente medida de calidad bayesiana: [16], [28].

$$P(G,D) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij}^{'})}{\Gamma(N_{ij}^{'}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}+N_{ijk}^{'})}{\Gamma(N_{ijk}^{'})} \tag{1.4}$$

donde:

$P(G)$: es la probabilidad a priori de la estructura G de la red. En caso de que ninguna información esté disponible se adopta una distribución uniforme.

D : describe el conjunto de datos.

n : el número de variables discretas.

r_i : describe la cantidad de valores posibles de la variable X_i .

q_i : es el número posible de configuraciones del conjunto $pa(X_i)$ (padres de X_i).

N_{ijk} : es el número de casos del conjunto de datos en que la variable X_i toma su k-ésimo valor y el conjunto $pa(X_i)$ toma su j-ésima configuración.

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$$

N'_{ijk} : son los hiperparámetros de la distribución de probabilidad Dirichlet.

$$N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$$

Medida de mínima longitud de descripción MDL

Esta medida caracteriza la estimación en términos de comprensión de los datos. El objetivo es encontrar un modelo que facilite la obtención de la descripción más corta posible de los datos originales. La longitud de esta descripción toma en cuenta:
- La descripción del propio modelo, penalizando la complejidad del mismo.
- La descripción de los datos que usa el modelo, alentando su verosimilitud.

Tanto la medida bayesiana como el MDL son bien conocidas y están bien estudiadas, ver [3], [12].

Métodos basados en detección de independencias

La estimación de la estructura de la red bayesiana por el método basado en detección de independencias consiste en, dado un conjunto de independencias condicionales en las variables de estudio, se intentará encontrar la estructura gráfica de la red la cual (la *condición de markov*) enlazará todas estas independencias condicionales.

Con este método se encuentra un único modelo basado en la información de las independencias condicionales obtenidas por pruebas estadísticas en los datos. Esto es, el método basado en pruebas usa un análisis estadístico para probar la presencia de independencia condicional. Un algoritmo para este método es el algoritmo PC [28].

1.2.2 Especificación de los parámetros de la red bayesiana

La especificación de los parámetros en la red bayesiana, llamada *estimación paramétrica*, consiste en especificar los parámetros de las distribuciones de probabilidad que se definen en la red.

Una vez conocida la estructura de la red, se procede a realizar la estimación paramétrica.

La estructura de la red determinará para cada variable X_i su conjunto de padres, denotado por $pa(X_i)$, y las correspondientes distribuciones de probabilidad condicionadas $P(x_i / pa(X_i))$ que se definen en la red.
Entonces el problema consiste en estimar los parámetros de las distribuciones de probabilidad condicionadas a partir de los datos.

La forma habitual y más simple de estimar las distribuciones de probabilidad condicional es mediante el cálculo de las frecuencias relativas de ocurrencia de los correspondientes sucesos. Esto es, el valor estimado de la probabilidad condicional viene determinada por el cociente entre el número de casos favorables y el de casos posibles:

$$P(x_i / pa(X_i)) = \frac{n(x_i, pa(X_i))}{n(Pa(X_i))} \tag{1.5}$$

donde:

$n(x_i, pa(X_i))$: es el número de casos de nuestro conjunto de datos en el que la variable X_i toma el valor x_i y que $pa(X_i)$ (padres de X_i) toma su configuración.

$n(pa(X_i))$: es el número de casos de nuestro conjunto de datos en que $pa(X_i)$ toma su configuración.

A pesar de que esta estimación aparenta ser una buena aproximación hay que tener en cuenta que, para problemas reales, suele suceder que los casos del conjunto de datos no abarcan todas las posibilidades de combinaciones entre valores de variables, con lo que este tipo de estimación puede llevar a una estimación de parámetros con abundancia de ceros.

Para atenuar este problema existen procedimientos de estimadores basados en suavizados. Uno de los más conocidos es el estimador basado en la *sucesión de Laplace* [15], que viene definido por la siguiente fórmula:

$$\frac{n(x_i, pa(X_i)) + 1}{n(Pa(X_i)) + r_i}$$

Ahora la estimación de la probabilidad viene expresada por el número de casos favorables +1 dividida por el de casos totales más el número de alternativas del atributo X_i denotado por r_{A_i} .

o equivalente,

$$\frac{N_{ijk}+1}{N_{ij}+r_i}$$

donde

N_{ijk} : es el número de casos del conjunto de datos en que la variable X_i toma su k-ésimo valor y el conjunto $pa(X_i)$ (padres de X_i) toma su j-ésima configuración.

N_{ij} : es el número de casos del conjunto de datos en que la variable X_i toma sus valores y el conjunto $pa(X_i)$ (padres de X_i) toma su j-ésima configuración.

r_i : es el número de valores posibles de la variable X_i .

La calidad de estas estimaciones dependerá de que exista un número suficiente de datos en la muestra. Cuando esto no es posible se puede cuantificar la incertidumbre existente representándola mediante una distribución de probabilidad, para así considerarla explícitamente en la definición de las probabilidades. Habitualmente se emplean distribuciones Beta en el caso de variables binarias, y Dirichlet para variables multivaluadas. Esta aproximación es útil cuando se encuentra con el apoyo de expertos en el dominio de la aplicación para concretar los valores de los parámetros de las distribuciones.

Si existen variables de tipo continuo la estrategia es discretrizarlas antes de construir el modelo estructural. Existen algunos modelos de redes bayesianas con variables continuas, pero están limitados a variables gaussianas relacionadas, ver [14].

1.3 Inferencia en la red bayesiana

Habiéndose construido una red bayesiana para un conjunto de variables aleatorias, ésta representa nuestro "modelo" para estas variables. Cuando se tiene información disponible de algunas de estas variables, podemos usar este "modelo" para realizar inferencia acerca de las variables no observables en la red.

La inferencia en redes bayesianas es realizado usando el Teorema de bayes. Considere una red para un conjunto de variables aleatorias $X = (X_1, X_2, ..., X_n)$ y asuma que alguna de las variables, $\boldsymbol{B}^*$, son observadas y el resto, $\boldsymbol{A}^*$, no lo son. Podemos entonces, usar el Teorema de bayes, para calcular la distribución condicional de $\boldsymbol{A}^*$ dado $\boldsymbol{B}^*$ como

$$P(A^* / B^*) \alpha P(B^* / A^*) P(A^*) \tag{1.6}$$

De este modo $\boldsymbol{P(A^*)}$ es la distribución a priori de $\boldsymbol{A}^*$, es decir, la distribución de $\boldsymbol{A}^*$ antes de observar $\boldsymbol{B}^*$, $\boldsymbol{P(B^* / A^*)}$ es la verosimilitud de $\boldsymbol{A}^*$ y $\boldsymbol{P(A^* / B^*)}$ es la distribución a posteriori de $\boldsymbol{A}^*$, es decir, la distribución de $\boldsymbol{A}^*$, cuando se ha observado $\boldsymbol{B}^*$.

Generalmente, encontrar estas distribuciones es computacionalmente pesado según los cálculos laboriosos que implica la distribución de probabilidad conjunta, especialmente si hay muchas variables en la red. Por lo tanto, métodos eficientes de implementación del Teorema de bayes son usados. Estas implementaciones usan la factorización de la distribución de la probabilidad conjunta de todas las variables en la red.

CAPÍTULO II

2. CLASIFICADOR NAIVE BAYES

Introducción

La tarea de clasificación busca clasificar o etiquetar correctamente un conjunto de datos en uno de los grupos o clases previamente definidas.

Un *clasificador* es una función que clasifica o asigna a un objeto o individuo en uno de los grupos o clases predefinidas.

Un *clasificador bayesiano* es una función que asigna a un objeto u observación en la clase con mayor probabilidad.

Dentro de los métodos de clasificación, los clasificadores bayesianos presentan un rendimiento bastante bueno y han demostrado ser tan competitivos como los árboles de decisión y redes neuronales [37], y se están usando exitosamente en muchas aplicaciones relacionadas con la clasificación [2], [11], [1].

Entre los clasificadores basados en redes bayesianas (Naive Bayes, Red Bayesiana aumentada a árbol TAN, Semi Naive Bayes y otros) el clasificador Naive Bayes ha demostrado comportarse bastante bien a pesar de que asume que las variables predictoras son condicionalmente independientes dada la clase, lo que generalmente no se cumple. Este clasificador Naive Bayes es uno de los modelos más efectivos, debido a su simplicidad, su resistencia al ruido, poco tiempo de procesamiento y alto poder predictivo [35], [24].

2.1 Clasificador bayesiano

La estructura de una red bayesiana puede utilizarse para construir clasificadores identificando uno de los nodos como la variable clase. El proceso de clasificación se realiza usando el teorema de Bayes para asignar a un objeto u observación en la clase con mayor probabilidad.

En el problema de clasificación, se tiene la variable clase ($\boldsymbol{C}$) y un conjunto de variables predictoras o atributos $\{A_1, A_2, ..., A_n\}$.

Para éstas variables se obtiene una red bayesiana $\boldsymbol{B}$, que define una distribución de probabilidad conjunta $P(C, A_1, A_2, ..., A_n)$. Entonces se puede utilizar la red bayesiana resultante para dado los valores de un conjunto de atributos $\{a_1, a_2, ..., a_n\}$, el clasificador basado en $\boldsymbol{B}$ retorna la etiqueta $\boldsymbol{c}$ que maximice la probabilidad a posteriori $P(C = c / a_1, a_2, ..., a_n)$, la clase **c** toma "m" posibles valores $c_1, c_2, ..., c_m$.

Un *clasificador bayesiano* se basa en el uso de una red bayesiana de las variables en estudio y a partir de ella obtiene el valor más probable de la variable (clase) dada cualquier configuración en el resto de variables (atributos).

El proceso de clasificación

Sea $A_1, A_2, ..., A_n$ las variables o atributos que permiten predecir el valor de la clase C. De acuerdo al teorema de bayes, la probabilidad de que una observación $\{a_1, a_2, ..., a_n\}$ pertenezca a la clase "$\boldsymbol{c}$" es,

$$P(C = c / a_1, a_2, ..., a_n) = \frac{P(C = c) P(a_1, a_2, ..., a_n / C = c)}{P(a_1, a_2, ..., a_n)} \qquad (2.1)$$

con $P(c, a_1, a_2, ..., a_n) = P(C = c) P(a_1, a_2, ..., a_n / C = c)$

El proceso de clasificación se realiza asignando la observación o conjunto de atributos $\{a_1, a_2, ..., a_n\}$ en aquella clase con mayor probabilidad.

Las redes bayesianas pueden ayudarnos a simplificar la representación de la función de probabilidad conjunta $P(c, a_1, a_2, ..., a_n)$ considerando las relaciones de dependencia que existen entre las variables.

2.2 Clasificador Naive Bayes

El clasificador Naive Bayes (llamado también clasificador bayesiano simple), es un clasificador bayesiano, esto es, es el que asigna a un objeto u observación en la clase con mayor probabilidad, y supone que todas las variables o atributos son condicionalmente independientes dado el valor de la clase.

Sea $A_1, A_2, ..., A_n$ las variables o atributos que permiten predecir el valor de la clase C.
La suposición de independencia asumida por el clasificador Naive Bayes da lugar a un modelo de red bayesiana con estructura simple descrita en la Figura 3.1 . En el existe un único nodo raíz (la clase), y en la que todos los atributos son nodos hoja que tienen como único padre a la variable clase.

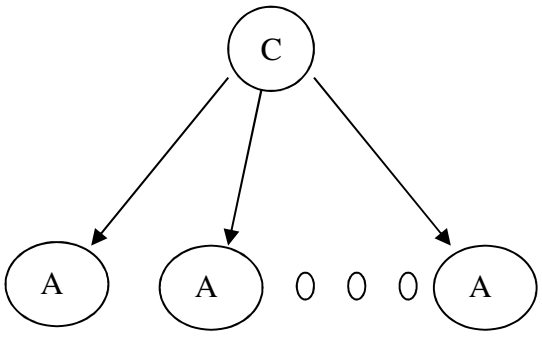


Figura 2.1.- Estructura del clasificador Naive Bayes

El *proceso de clasificación* Naive Bayes utiliza la anterior ecuación (2.1). Tomando en cuenta la suposición de que los atributos son condicionalmente independientes dada la clase, la probabilidad de que una observación o conjunto de atributos $\{a_1, a_2, ..., a_n\}$ pertenezca a la clase "***c***" es

$$P(C = c / a_1, a_2, ..., a_n) = \frac{P(C = c)P(a_1 / C = c)P(a_2 / C = c)...P(a_n / C = c)}{P(a_1, a_2, ..., a_n)} \qquad (2.2)$$

factorizando

$$P(C = c / a_1, a_2, ..., a_n) = \frac{P(C = c)\prod_{i=1}^{n} P(a_i / C = c)}{P(a_1, a_2, ..., a_n)} \qquad (2.3)$$

con $c = c_1, c_2, ..., c_m$

esta ecuación se utilizará para la obtención de las probabilidades en la tarea de clasificación.

2.3 Especificación de los parámetros de la red del clasificador Naive Bayes

La especificación de los parámetros de la red Naive Bayes consiste en estimar los parámetros de las distribuciones de probabilidad que se definen en la red.

Esto es, definida la red Naive Bayes se tendrá que estimar los parámetros de la distribución de probabilidad de la variable clase $P(C)$ y de las distribuciones condicionales de los atributos dada la clase $P(A_i / C)$, estas distribuciones son representadas en la Figura 2.2 .

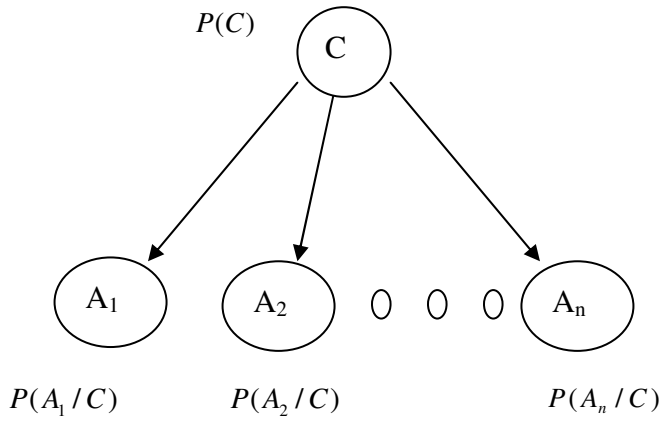


Figura 2.2.– Red del clasificador Naive Bayes

La estimación de los parámetros va a depender de que las variables o atributos A_i sean discretos o continuos.

Para atributos discretos

La estimación de la distribución de probabilidad de la clase $P(C)$ y de la distribución condicional $P(A_i / C)$ se basa en la frecuencia relativa de ocurrencia de los correspondientes sucesos, que se determinan en el conjunto de datos. Entonces el valor estimado de probabilidad es:

$$P(a_i / C = c) = \frac{n(a_i, C = c)}{n(C = c)} \tag{2.4}$$

donde,

$n(a_i, C = c)$: es el número de casos del conjunto de datos en que la variable A_i toma el valor a_i y su padre (clase) $\boldsymbol{C}$ toma el valor de $\boldsymbol{c}$.

$n(C = c)$: es el número de casos del conjunto de datos en que la clase $\boldsymbol{C}$ toma el valor de $\boldsymbol{c}$.

A pesar de que esta estimación aparenta ser una buena aproximación, para problemas reales, suele suceder que los casos del conjunto de datos no abarcan todas las combinaciones de valores de la variable clase con las variables predictoras, con lo que este tipo de estimación puede llevar a una estimación de parámetros con abundancia de ceros.

Existen varios métodos que intentan solucionar estos problemas. Una de ellos es el estimador basado en la *ley de sucesión de Laplace* [15], en la cual, en lugar de estimar la probabilidad directamente como,

$$\frac{\boldsymbol{Casos\ \ favorables}}{\boldsymbol{Casos\ \ Totales}},$$

utiliza como estimación el número que se obtiene al dividir

$$\frac{n(a_i, C = c) + 1}{n(C = c) + r_{A_i}}$$

es decir, el número de casos favorables más uno dividido por el número de casos totales más el número de valores posibles del atributo A_i denotado por r_{A_i} .

o equivalente,

$$\frac{N_{ik} + 1}{N_i + r_i}$$

donde

N_{ik} : es el número de casos del conjunto de datos en que la variable A_i toma su k-ésimo valor y la clase C toma su valor c.

N_i : es el número de casos del conjunto de datos en que la variable A_i toma valores y la clase C toma su valor c .

r_i : es el número de valores posibles de la variable A_i .

Para atributos continuos

El clasificador Naïve Bayes puede ser aplicado también cuando hay variables predictoras continuas, hay dos alternativas:

1. Aplicar previamente un método de discretización de la variable continua.
2. Asumiendo una distribución para cada variable predictora, por lo general gaussiana, con media y varianza estimada de los datos.

En el segundo caso, el clasificador Naive Bayes supone que la variable predictora en cuestión sigue una distribución normal; por tanto, lo único que se calcula de los datos es la media μ y la desviación típica σ condicionadas a cada valor de la variable clase.

$$P(A_i / c) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(X - \mu)^2}{2\sigma^2} \right) \tag{2.5}$$

Evidentemente, esta estimación tiene el inconveniente de que los datos no siempre siguen una distribución normal.

La presente tesis se centra en el uso de variables discretas, por lo que siempre que se tenga variables continuas estas serán discretizadas antes de su uso.

Discretización de variables continuas

En general hay tres razones principales que han llevado a prestar gran atención a la discretización de variables continuas [9]:

- Muchos algoritmos desarrollados en aprendizaje automático sólo son capaces de aprender desde variables discretas. Debido a que muchas tareas de clasificación de la vida real presentan valores continuos, se hace necesaria una discretización para poder utilizar estos algoritmos.

- Algunos clasificadores, como ocurre con Naíve-Bayes, pueden ser utilizados tanto sobre datos discretos como sobre datos continuos, aunque la correcta discretización de los datos puede llevar a una mejora en su rendimiento.

- Mejora de velocidad en los algoritmos de inducción que utilizan atributos discretos.

Los métodos de discretización son desarrollados: usando intervalos de igual ancho, usando intervalos con igual frecuencia, ChiMerge, 1R, discretización usando el método de la entropía, etc.

En los experimentos se emplea la discretización usando intervalos con igual frecuencia con un valor de diez para el número de intervalos, ya que demostró buenos resultados [22], [5], [39], estos estudios fueron realizados en grupos de datos reales de diferentes dominios.

El método de discretización especificado fue implementado con el programa WEKA [38].

2.4 Caso ilustrativo

DIAGNÓSTICO DE USO DE LENTES DE CONTACTO

Para el diagnóstico de lentes de contacto se toman como base atributos relevantes como la edad, padecimiento, astigmatismo y lagrimeo. De acuerdo con estos datos se puede predecir el uso o no de lentes y de qué tipo.

Para mostrar la aplicación del clasificador Naive Bayes se utilizó un conjunto de datos que proviene del artículo de [7], que contiene 24 casos, en el cual se representan las características de los pacientes, que se muestra en el siguiente Cuadro 2.1 . A cada caso le corresponde el diagnóstico de no usar lentes o de usar lentes duros o suaves. El diagnóstico, es la clase que se quiere predecir a partir las características de los pacientes, edad, padecimiento, si sufre de astigmatismo y lagrimeo.

Cuadro 2.1.- Conjunto de datos del Diagnóstico de lentes de contacto

Paciente	Edad	Padecimiento	Astigmatismo	Lagrimeo	Tipo de lente
1	Joven	Hipermétrope	Si	reducido	Ninguno
2	Joven	Hipermétrope	Si	normal	Duro
3	Joven	Hipermétrope	No	reducido	Ninguno
4	Joven	Hipermétrope	No	normal	Suave
5	Joven	Miope	Si	reducido	Ninguno
6	Joven	Miope	Si	normal	Duro
7	Joven	Miope	No	reducido	Ninguno
8	Joven	Miope	No	normal	Suave
9	pre-presbiópico	Hipermétrope	Si	reducido	Ninguno
10	pre-presbiópico	Hipermétrope	Si	normal	Ninguno
11	pre-presbiópico	Hipermétrope	No	reducido	Ninguno
12	pre-presbiópico	Hipermétrope	No	normal	Suave
13	pre-presbiópico	Miope	Si	reducido	Ninguno
14	pre-presbiópico	Miope	Si	normal	Duro
15	pre-presbiópico	Miope	No	reducido	Ninguno
16	pre-presbiópico	Miope	No	normal	Suave
17	Presbiópico	Hipermétrope	Si	reducido	Ninguno
18	Presbiópico	Hipermétrope	Si	normal	Ninguno
19	Presbiópico	Hipermétrope	No	reducido	Ninguno
20	Presbiópico	Hipermétrope	No	normal	Suave
21	Presbiópico	Miope	Si	reducido	Ninguno
22	Presbiópico	Miope	Si	normal	Duro
23	Presbiópico	Miope	No	reducido	Ninguno
24	Presbiópico	Miope	No	normal	Ninguno

Las variables y sus posibles valores se presentan en el siguiente cuadro,

Cuadro 2.**2**.- Descripción de las variables del conjunto de datos Diagnóstico de lentes de contacto

Variable	**Valores**	**Notación**
A_1 : Edad	Joven, Pre-presbiópico, Presbiópico	$A_1 : a_{11}, a_{12}, a_{13}$
A_2 : Padecimiento	Hipermétrope, Miope	$A_2 : a_{21}, a_{22}$
A_3 : Astigmatismo	Si, No	$A_3 : a_{31}, a_{32}$
A_4 : Lagrimeo	Normal, Reducido	$A_4 : a_{41}, a_{42}$
C : *Clase (Tipo de lente)*	*Ninguno, Suave, Duro*	$C : c_1, c_2, c_3$

En el cuadro de datos, el caso 7 muestra a un individuo joven que padece de miopía, no sufre astigmatismo y presenta lagrimeo reducido, se le diagnostica como no apto para uso de lentes de contacto.

La red del clasificador Naive Bayes para los atributos A_1, A_2, A_3, A_4 y la clase C esta descrita en la Figura 2.3.

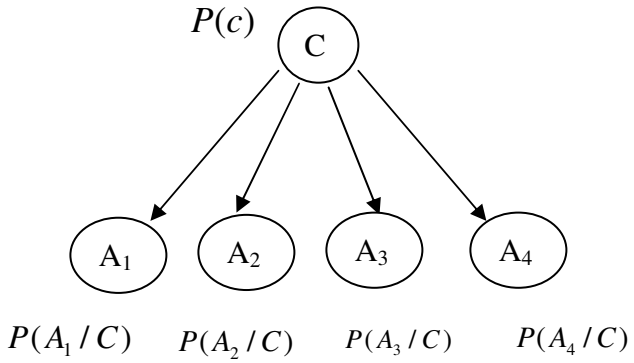


Figura 2.3.- Red del clasificador Naive Bayes para el diagnóstico del tipo de lente

En el desarrollo del proceso de clasificación, se tendrá que estimar los parámetros de la distribución de probabilidad de la variable clase $P(C)$ y de las distribuciones condicionales de los atributos dada la clase $P(A_i / C)$ que se definen en el clasificador Naive Bayes. Se recomienda para agilizar el análisis los siguientes pasos:

1) Organizar los datos

La forma de organizar los datos es a través de un conteo en los casos, los cuales son mostrados en las tablas siguientes:

Tablas de contingencia o conteo

	Tipo de lente (C)		
Edad (A_1)	Ninguno	Suave	Duro
Joven	4	2	2
Pre-presbiópico	5	2	1
Presbiópico	6	1	1

	Tipo de lente (C)		
Padecimiento (A_2)	Ninguno	Suave	Duro
Hipermétrope	8	3	1
Miope	7	2	3

	Tipo de lente (C)		
Astigmatismo (A_3)	Ninguno	Suave	Duro
Si	8	0	4
No	7	5	0

	Tipo de lente (C)		
Lagrimeo (A_4)	Ninguno	Suave	Duro
Normal	3	5	4
Reducido	12	0	0

Tipo de lente (C)		
Ninguno	Suave	Duro
15	5	4

2) Cálculo de las distribuciones de probabilidad **(Tabla de probabilidades)**

Para la obtención de la probabilidad marginal y las probabilidades condicionales se empleó la *corrección de Laplace*, ya que los casos del conjunto de datos no abarcaron todas las combinaciones de valores de la variable clase con las variables predictoras (ver tabla de conteo anterior).

El programa Weka, en su aplicación del clasificador Naive Bayes emplea la *corrección de Laplace* para el cálculo de las probabilidades respectivas.

- Distribución de probabilidad de la Clase $\boldsymbol{C}$,

C	C_1	C_2	C_3
$\boldsymbol{P(C)}$	16/27	6/27	5/27

Empleando la corrección de Laplace que se definió en la sección anterior y de la tabla de conteo, se tiene:

$$P(c_1)=\frac{15+1}{24+3}=\frac{16}{27} \qquad P(c_2)=\frac{5+1}{24+3}=\frac{6}{27} \qquad P(c_3)=\frac{4+1}{24+3}=\frac{5}{27}$$

- Tablas de las distribuciones de probabilidad condicional, esto es, $\boldsymbol{P(A_i/C)}$,

$\boldsymbol{C}$	$\boldsymbol{P(A_1/C)}$		
C_1	5/18	6/18	7/18
C_2	3/8	3/8	2/8
C_3	3/7	2/7	2/7

$\boldsymbol{C}$	$\boldsymbol{P(A_2/C)}$	
c_1	9/17	8/17
c_2	4/7	3/7
c_3	2/6	4/6

$\boldsymbol{C}$	$\boldsymbol{P(A_3/C)}$	
c_1	9/17	8/17
c_2	1/7	6/7
c_3	5/6	1/6

$\boldsymbol{C}$	$\boldsymbol{P(A_4/C)}$	
c_1	4/17	13/17
c_2	6/7	1/7
c_3	5/6	1/6

Al considerar el caso de padecer astigmatismo dado que se usa lente suave (en la tabla de conteo). La probabilidad asignada entonces es de 0/5, lo cual eliminaría la posibilidad

de que se presente este caso, pero en circunstancias reales no se puede asegurar a priori que nunca se presentará al consultorio una persona así.
Por ello, para eliminar el problema presentado, se utilizó la *corrección de Laplace*.

Esto es, para hallar la probabilidad condicional se utilizó la expresión

$$\frac{n(a_i, C = c) + 1}{n(C = c) + r_i}$$

es decir, el número de casos favorables más uno dividido por el número de casos totales más el (r_i) número de valores posibles del atributo A_i .

Entonces, la probabilidad condicional de padecer astigmatismo dado que se usa lente suave, el estimador de Laplace es:

$$P(a_{31} / c_2) = \frac{0+1}{5+2} = \frac{1}{7}$$

y la probabilidad de no padecer astigmatismo dado que usa lente suave es:

$$P(a_{32} / c_2) = \frac{5+1}{5+2} = \frac{6}{7}$$

Se indica que, en el cálculo de todas las probabilidades se empleó la *corrección de Laplace*.

En el caso de obtener la probabilidad de la Edad dado el resultado del tipo de lente ninguno, el resultado será:

$$P(a_{11} / c_1) = \frac{4+1}{15+3} = \frac{5}{18} \qquad P(a_{12} / c_1) = \frac{5+1}{15+3} = \frac{6}{18} \qquad P(a_{13} / c_1) = \frac{6+1}{15+3} = \frac{7}{18}$$

3) Clasificación

Para los datos

Usando la ecuación que se definió para la clasificación, esto es,

$$P(C = c / a_1, a_2, ..., a_n) = \frac{P(C = c)\prod_{i=1}^{n} P(a_i / C = c)}{P(a_1, a_2, ..., a_n)} \quad (2.6)$$

los resultados obtenidos por el clasificador se muestran en la última columna del Cuadro 2.3, que muestra también las características de los pacientes

Cuadro 2.3 .- Conjunto de datos del Diagnóstico de lente de contacto y la probabilidad de predicción

Paciente	Edad	Padecimiento	Astigmatismo	Lagrimeo	Tipo de lente	*Probabilidad*
1	joven	Hipermétrope	Si	reducido	Ninguno	**0.884**
2	joven	Hipermétrope	Si	normal	Duro	**0.524**
3	joven	Hipermétrope	No	reducido	Ninguno	**0.827**
4	joven	Hipermétrope	No	normal	Suave	**0.724**
5	joven	Miope	Si	reducido	Ninguno	**0.795**
6	joven	Miope	Si	normal	Duro	**0.724**
7	joven	Miope	No	reducido	Ninguno	**0.827**
8	joven	Miope	No	normal	Suave	**0.622**
9	pre-presbiópico	Hipermétrope	Si	reducido	Ninguno	**0.925**
10	pre-presbiópico	Hipermétrope	Si	normal	Ninguno	**0.419**
11	pre-presbiópico	Hipermétrope	No	reducido	Ninguno	**0.856**
12	pre-presbiópico	Hipermétrope	No	normal	Suave	**0.714**
13	pre-presbiópico	Miope	Si	reducido	Ninguno	**0.870**
14	pre-presbiópico	Miope	Si	normal	Duro	**0.606**
15	pre-presbiópico	Miope	No	reducido	Ninguno	**0.862**
16	pre-presbiópico	Miope	No	normal	Suave	**0.633**
17	presbiópico	Hipermétrope	Si	reducido	Ninguno	**0.941**
18	presbiópico	Hipermétrope	Si	normal	Ninguno	**0.485**
19	presbiópico	Hipermétrope	No	reducido	Ninguno	**0.909**
20	presbiópico	Hipermétrope	No	normal	Suave	**0.594**
21	presbiópico	Miope	Si	reducido	Ninguno	**0.891**
22	presbiópico	Miope	Si	normal	Duro	**0.599**
23	presbiópico	Miope	No	reducido	Ninguno	**0.909**
24	presbiópico	Miope	No	normal	Ninguno	**0.349**

Para obtener la primera probabilidad, se tomará los datos del primer paciente,

Paciente	Edad	Padecimiento	Astigmatismo	Lagrimeo	Tipo de lente	*Probabilidad*
1	Joven	Hipermétrope	Si	reducido	Ninguno	

entonces la probabilidad para esta combinación de atributos será: (reemplazando en la ecuación anterior)

$$P(C=c/a_{11},a_{21},a_{31},a_{42})=\frac{P(C=c)P(a_{11}/C=c)P(a_{21}/C=c)P(a_{31}/C=c)P(a_{42}/C=c)}{P(a_{11},a_{21},a_{31},a_{42})}$$

$$P(C=c/a_{11},a_{21},a_{31},a_{42})=\alpha P(C=c)P(a_{11}/C=c)P(a_{21}/C=c)P(a_{31}/C=c)P(a_{42}/C=c)$$

donde α es la constante de proporcionalidad

luego, se tendrá que hallar las probabilidades para los valores de la clase $C:c_1,c_2,c_3$

$$P(c_1/a_{11},a_{21},a_{31},a_{42})=\alpha P(c_1)P(a_{11}/c_1)P(a_{21}/c_1)P(a_{31}/c_1)P(a_{42}/c_1)$$

$$P(c_2/a_{11},a_{21},a_{31},a_{42})=\alpha P(c_2)P(a_{11}/c_2)P(a_{21}/c_2)P(a_{31}/c_2)P(a_{42}/c_2)$$

$$P(c_3/a_{11},a_{21},a_{31},a_{42})=\alpha P(c_3)P(a_{11}/c_3)P(a_{21}/c_3)P(a_{31}/c_3)P(a_{42}/c_3)$$

reemplazando, de las tablas de probabilidades

$$P(c_1/a_{11},a_{21},a_{31},a_{42})=\alpha\frac{16}{27}\frac{5}{18}\frac{9}{17}\frac{9}{17}\frac{13}{17}=\alpha(0.03528)$$

$$P(c_2/a_{11},a_{21},a_{31},a_{42})=\alpha\frac{6}{27}\frac{3}{8}\frac{4}{7}\frac{1}{7}\frac{1}{7}=\alpha(0.00097)$$

$$P(c_3 / a_{11}, a_{21}, a_{31}, a_{42}) = \alpha \frac{5}{27}\frac{3}{7}\frac{2}{6}\frac{5}{6}\frac{1}{6} = \alpha(0.00367)$$

normalizando los valores anteriores, obtenemos

$$\alpha(0.03528 + 0.00097 + 0.00367) = 1 \qquad \alpha = 25.05$$

$$P(c_1 / a_{11}, a_{21}, a_{31}, a_{42}) = \alpha(0.03528) = 0.884$$

$$P(c_2 / a_{11}, a_{21}, a_{31}, a_{42}) = \alpha(0.00097) = 0.024$$

$$P(c_3 / a_{11}, a_{21}, a_{31}, a_{42}) = \alpha(0.00367) = 0.092$$

Luego, P(Ninguno / las evidencias) = 0.884
P(Lente suave / evidencias) = 0.024
P(Lente duro / evidencias) = 0.092

Para nuevos datos

Para ilustrar la aplicación práctica del método, se supone que se presenta un nuevo paciente con las características siguientes:

Edad	Padecimiento	Astigmatismo	Lagrimeo	Tipo de lente	***Probabilidad***
Joven	hipermétrope	No	Normal	……..	

Para predecir el tipo de lente recomendado para un paciente con estas características, se utilizará las probabilidades condicionales en la ecuación para la clasificación.

entonces reemplazando,

$$P(C=c/a_{11},a_{21},a_{32},a_{41})=\frac{P(C=c)P(a_{11}/C=c)P(a_{21}/C=c)P(a_{32}/C=c)P(a_{41}/C=c)}{P(a_{11},a_{21},a_{32},a_{41})}$$

$$P(C=c/a_{11},a_{21},a_{32},a_{41})=\alpha P(C=c)P(a_{11}/C=c)P(a_{21}/C=c)P(a_{32}/C=c)P(a_{41}/C=c)$$

luego, se tendrá que hallar las probabilidades para los valores de la clase $C:c_1,c_2,c_3$

$$P(c_1/a_{11},a_{21},a_{32},a_{41})=\alpha\frac{16}{27}\frac{5}{18}\frac{9}{17}\frac{8}{17}\frac{4}{17}=\alpha(0.00965)$$

$$P(c_2/a_{11},a_{21},a_{32},a_{41})=\alpha\frac{6}{27}\frac{3}{8}\frac{4}{7}\frac{6}{7}\frac{6}{7}=\alpha(0.03499)$$

$$P(c_3/a_{11},a_{21},a_{32},a_{41})=\alpha\frac{5}{27}\frac{3}{7}\frac{2}{6}\frac{1}{6}\frac{5}{6}=\alpha(0.00367)$$

normalizando

$$\alpha(0.00965+0.03499+0.00367)=1 \quad , \quad \alpha=20.7$$

$$P(c_1/a_{11},a_{21},a_{32},a_{41})=\alpha(0.00965)=0.200$$

$$P(c_2/a_{11},a_{21},a_{32},a_{41})=\alpha(0.03499)=0.724$$

$$P(c_3/a_{11},a_{21},a_{32},a_{41})=\alpha(0.00367)=0.076$$

Luego, P(Ninguno / las evidencias) = 0.200
P(Lente suave / evidencias) = 0.724
P(Lente duro / evidencias) = 0.076

Con este método se obtienen las probabilidades de diagnóstico para cada tipo de lente, dependiendo de la combinación de los atributos del caso a clasificar.

El clasificador bayesiano asigna a un individuo a la clase con mayor probabilidad a posteriori, por ello, para un paciente con las características siguientes de ser joven, hipermétrope, no padecer astigmatismo y presentar lagrimeo normal, será diagnosticado con el uso de lente suave, ya que se tiene una mayor probabilidad (0.724).
El clasificador Naive Bayes tiene una ventaja importante, que además de realizar la clasificación esta viene acompañada con un valor de probabilidad de ocurrencia, lo que serviría para analizar el nivel de certeza que se tendría en la clasificación.

CAPÍTULO III

ÁRBOLES DE DECISIÓN

Introducción

Los árboles de decisión se constituyen como uno de los modelos más utilizados en tareas de clasificación [33]. El conocimiento obtenido durante el proceso de construcción del modelo se representa mediante un árbol en el cual cada nodo interno contiene un atributo particular (con un nodo hijo para cada posible valor del atributo) y en el que cada hoja se refiere a una decisión (etiquetada con una de las clases del problema).

Un árbol de decisión puede usarse para clasificar un caso comenzando de su raíz y siguiendo el camino determinado por los valores de los atributos en los nodos internos hasta que encontremos una hoja del árbol que se refiere a la decisión (clasificación).

Existe una serie de algoritmos desarrollados desde los principios de los 60's para la construcción de árboles de decisión. CLS ([17]Hunt et al., 1966) para modelar aprendizaje humano de conceptos, ID3 ([32]Quinlan, 1986) desarrollado con el criterio de ganancia de información para desarrollar sistemas expertos desde ejemplos (casos preclasificados), CART ([4]Breiman et al., 1984) (Classification And Regression Tree, Arboles de clasificación y regresión) es un sistema recursivo binario de particiones, CHAID ([20]Kass, 1980) (Chi Square Automatic Interaction Detection, Detección de Interacción Automática de Chi Cuadrado) que es un algoritmo recursivo de clasificación no binario, C4.5 ([33]Quinlan, 1993) en él se actualiza (mejora) los sistemas de decisión, etc.

Entre los algoritmos para construir árboles de decisión, el ID3 y el sucesor C4.5, son considerados los más populares para la tarea de clasificación.

Los algoritmos de construcción de árboles de decisión suelen construir de forma descendente los árboles de decisión, comenzando en la raíz del árbol. Por este motivo se suele hacer referencia a este tipo de algoritmos como pertenecientes a la familia TDIDT (Top-Down Induction of Decisión Trees, Inducción Descendente de los Mejores Arboles de Decisión).
La familia de algoritmos TDIDT abarca los algoritmos ID3, C4.5 .
La metodología para la clasificación, utilizando un árbol de decisión, se puede resumir en dos etapas: la primera es la etapa de construcción del árbol a partir de un conjunto de casos, o ejemplos (datos de entrenamiento) y la segunda es la etapa de clasificación, donde nuevos casos pueden ser clasificados por el árbol construido.

Un *árbol de decisión* es una estructura gráfica compuesto por nodos (internos y hojas) y de arcos (ramas). Cada nodo interno está caracterizado por un atributo o variable (con u nodo hijo para cada posible valor del atributo) y en el que cada hoja se refiere a una decisión (etiquetada con una de las clases del problema).

Las características principales de un árbol de decisión son: su construcción sencilla, no necesita determinar de antemano parámetros para su construcción, puede tratar problemas multi-clase de la misma forma en que trabaja con problemas de dos clases y la fácil interpretación de su estructura.

En la Figura 4.1 se muestra un ejemplo de árbol de decisión que se puede aplicar a un nuevo paciente para saber si se le ha de recomendar o no una operación de cirugía ocular. Para esta tarea, basta recorrer a partir del nodo raíz (Astigmatismo) y seguir los valores respectivos de los atributos (del paciente) hasta alguna de las hojas del árbol, catalogadas como un “no” o “si”.

Este árbol de decisión en concreto funciona como un "*clasificador*", es decir, dado un nuevo individuo lo clasifica en una de las dos clases posibles: "no" o "si".

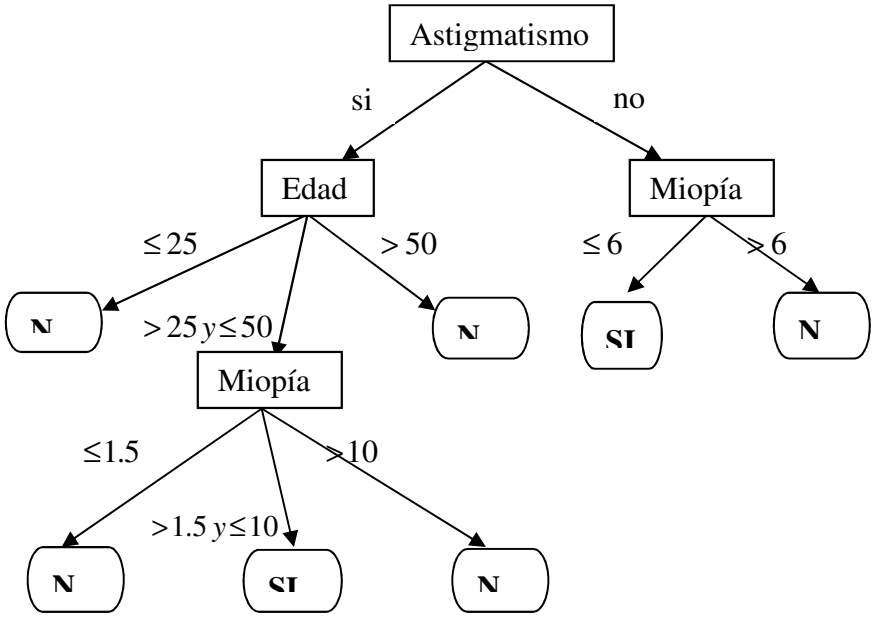


Figura 3.1.– Árbol de decisión para recomendar o no una cirugía ocular

3.1 Construcción de los árboles de decisión

Los árboles de decisión se construyen a partir del método basado en un particionamiento recursivo del conjunto de casos [17].

La estructura de este método para construir un árbol de decisión a partir de un conjunto T de casos se detalla a continuación.

Sea T el conjunto de casos ya clasificados (datos de entrenamiento) con "n" atributos o variables predictoras $\{A_1, A_2, ..., A_n\}$ y una variable clase C con valores c_j, $j = 1,2,...,k$.

Existen tres posibilidades:

1. Si T contiene uno o más casos, y todos pertenecientes a una única clase c_j, el árbol de decisión para T es una hoja etiquetada con la clase c_j.

2. Si T no contiene ningún caso, el árbol de decisión es una hoja, donde la clase asociada a dicha hoja debe ser determinada por información que no pertenece a T. Por ejemplo, la hoja puede etiquetarse de acuerdo a conocimientos base del dominio (conocimiento del problema), como ser la clase mayoritaria.

3. Si T contiene casos pertenecientes a varias clases,

   La idea es refinar (particionar) el conjunto de casos T en subconjuntos que tiendan a contener casos pertenecientes a una única clase.

   Para ello, se selecciona un atributo (o variable) A, que tiene "m" posibles valores y T se particiona en los subconjuntos $T_1, T_2, ..., T_m$ donde T_i contiene todos los casos de T que tienen el valor *i-ésimo* del atributo seleccionado.

   El árbol de decisión para T consiste en que cada nodo interno contiene un atributo (o

variable) con una rama para cada valor posible del atributo. El mecanismo de construcción del árbol se aplica recursivamente a cada subconjunto de casos.

En el proceso de *construcción del árbol de decisión* , en cada nodo interno se selecciona un atributo (o variable) que mejor particione o discrimine el conjunto de casos (subconjuntos que tiendan a contener casos pertenecientes a una única clase, o que consigan nodos más puros) con respecto de un criterio previamente establecido, *criterio de partición*.(el C4.5 y su predecesor el ID3, usan fórmulas basadas en teoría de la información para evaluar la bondad del atributo). El mecanismo de construcción del árbol se aplica recursivamente a cada subconjunto de casos, hasta una *condición de parada*.

3.1.1 Criterio de partición

Lo que se busca es seleccionar el mejor atributo (o variable) que particione al conjunto de casos en el proceso de construcción del árbol de decisión. Para esto, los criterios más conocidos son la ganancia de información usada por ID3, el criterio de proporción de la ganancia de información de C4.5 o el índice de diversidad de Gini empleado en CART.

Los criterios de partición o división utilizados generalmente se basan en la teoría de la información, esto es, basados en medidas de la impureza de un nodo, específicamente la *entropía* (la cual determina la impureza o incertidumbre de un conjunto de datos).
La bondad de una partición es la disminución de la impureza que se consigue con ella. Lo que se quiere es buscar particiones que discriminen o que consigan nodos más puros.

Se presenta el criterio de ganancia de información y la proporción de ganancia de información.

Criterio de ganancia de información

El criterio de ganancia de información busca seleccionar el atributo con mayor ganancia de información.

Sea un conjunto de casos $\boldsymbol{T}$, con múltiples clases $\boldsymbol{c_j}$, $j = 1,2,...,k$, la *entropía* de $\boldsymbol{T}$ se mide como:

$$I(T) = -\sum_{j=1}^{k} p_j \log_2 p_j \tag{3.1}$$

donde p_j es la proporción de casos en $\boldsymbol{T}$ que pertenecen a la clase c_j (o probabilidad de que un caso tomado al azar de $\boldsymbol{T}$ pertenezca a la clase c_j).

Esto es, $I(T)$ es la cantidad de información necesaria para clasificar un caso del conjunto de casos $\boldsymbol{T}$.

Si el atributo (o variable) $\boldsymbol{A}$ divide el conjunto $\boldsymbol{T}$ en los subconjuntos T_i , $i = 1,2,...,m$, entonces, la entropía total del sistema de subconjuntos será:

$$I(T, A) = \sum_{i=1}^{m} P(T_i) I(T_i) \tag{3.2}$$

donde $I(T_i)$ es la entropía del subconjunto T_i y $P(T_i)$ es la proporción de casos en $\boldsymbol{T}$ que pertenecen a $\boldsymbol{T_i}$, pueden calcularse, utilizando los tamaños relativos, $n(T_i)$, de los subconjuntos, como:

$$P(T_i) = \frac{n(T_i)}{n(T)}$$

Usando la entropía, la *ganancia de información*, de un atributo $\boldsymbol{A}$ en un conjunto de casos $\boldsymbol{T}$, se define como:

$$Ganancia(T, A) = I(T) - I(T, A) \tag{3.3}$$

donde $I(T)$ es el valor de la entropía a priori antes de realizar la subdivisión y $I(T, A)$ es el valor de la entropía del sistema de subconjuntos generados por la partición según $\boldsymbol{A}$.

Así, el atributo $\boldsymbol{A}$ seleccionado para determinar la división, será aquel que mayor ganancia obtenga respecto al conjunto $\boldsymbol{T}$, y se obtendrá a partir de (3.3).

El criterio de ganancia de información tiene un defecto muy serio y es que presenta una tendencia muy fuerte a favorecer atributos (o variables) con muchos valores (resultados).

Criterio de proporción de ganancia

Esta medida tiene en cuenta tanto la ganancia de información como las probabilidades de los distintos valores del atributo. Dichas probabilidades son recogidas mediante la denominada información de separación (split information), que no es más que la entropía del conjunto de datos $\boldsymbol{T}$ respecto a los valores del atributo A_i en consideración, siendo calculada como

$$I\_separación(A) = -\sum_{i=1}^{m} P(T_i) x \log_2 P(T_i) \tag{3.4}$$

que representa la información particionada generada al dividir $\boldsymbol{T}$ en "$\boldsymbol{m}$" subconjuntos.

La información de separación simboliza la información potencial que representa dividir el conjunto de datos, y es usada para compensar la menor ganancia de aquellos test (variables) con pocas salidas (valores). Con ello, tal y como se muestra en (3.5), la proporción de ganancia es calculada como el cociente entre la ganancia de información (3.3) y la información de separación (3.4). Tal cociente expresa la proporción de información útil generada por la división.

$$\text{Pr}oporción\ \ de\ \ ganancia(A) = \frac{Ganancia(T,A)}{I\_separación(A)} = \frac{I(T) - I(T,A)}{I\_separación(A)} \tag{3.5}$$

C4.5 maximiza este criterio de separación, premiando así a aquellos atributos que, aun teniendo una ganancia de información menor, disponen también de menor número de valores para llevar a cabo la clasificación. Sin embargo, si el test incluye pocos valores, la información de separación puede ser cercana a cero, y por tanto el cociente sería

inestable. Para evitar tal situación, el criterio selecciona un test que maximice la razón de ganancia pero obligando a que la ganancia del mismo sea al menos igual a la ganancia media de todos los test examinados. C4.5 ha resultado ser un sistema muy efectivo en la práctica, capaz de ofrecer una representación relativamente simple de los resultados con un bajo coste computacional.

3.1.2 Condición de parada

Cuando se detiene la construcción del árbol de decisión, se construye una hoja a la que se le puede asignar la clase más común de las recogidas por los casos.
Las reglas de parada tratan de predecir si merece la pena seguir construyendo el árbol o no. Ejemplos de este tipo de reglas son: pureza del nodo, cota de profundidad, mínimo de casos.

3.2 Árbol de decisión C4.5

El algoritmo del árbol de decisión C4.5 fue propuesto por Quinlan a finales de los años 80 para mejorar las carencias de su predecesor ID3. Desde entonces, ha sido uno de los sistemas clasificadores más referenciados en la bibliografía, principalmente debido a su extremada robustez en un gran número de dominios y su bajo coste computacional.
C4.5 introduce principalmente las siguientes mejoras:

1. Trata eficazmente los valores desconocidos.
2. Maneja los atributos continuos, aplicando una discretización previa.
3. Corrige la tendencia de ID3 a seleccionar los atributos con muchos valores distintos para establecer los test cambiando el criterio de división.

Este algoritmo genera árboles de decisión a partir de ejemplos mediante particiones realizadas recursivamente.

En cada nodo, el algoritmo debe decidir cual atributo o variable elegir para particionar los datos.

Los tipos de pruebas propuestos por Quinlan para C4.5 [33] son:

- Si el atributo es discreto, la representación es con un resultado y una rama para cada valor posible de la variable.
- Si el atributo es continuo, se realiza una prueba binaria con resultados A<= Z y A > Z, que comparan el valor de A con el umbral Z. Para calcular Z, se aplica un método, el cual ordena el conjunto de ***t*** valores distintos del atributo A presentes en el conjunto de entrenamiento, obteniendo el conjunto de valores $\{a_{i1}, a_{i2}, ..., a_{it}\}$. Cada par de valores consecutivos aporta un posible umbral $Z = \frac{a_{iv} + a_{i(v+1)}}{2}$, teniendo en total t-1 umbrales, donde t es como máximo el número de ejemplos. Una vez calculados los umbrales, C4.5 selecciona aquel que maximiza el criterio de separación (proporción de ganancia).

El algoritmo C4.5 incorpora una poda del árbol de decisión una vez que este ha sido inducido. La poda está basada en la aplicación de un test de hipótesis que trata de responder a la pregunta de si merece la pena expandir o no una determinada rama.

CAPÍTULO IV

OPTIMIZACIÓN DEL CLASIFICADOR NAIVE BAYES

En este trabajo, se presenta un método que busca optimizar el clasificador Naive Bayes usando el árbol de decisión C4.5 .

El propósito fue maximizar el rendimiento del clasificador Naive Bayes removiendo atributos redundantes y/o irrelevantes del conjunto de datos, y escoger los que son más informativos en tareas de clasificación, en esta labor se usó el árbol de decisión construido por el algoritmo C4.5 .

Lo que se hizo fue integrar al proceso de estimación del clasificador Naive Bayes, un proceso previo de selección de variables. En él, se eligió a partir de todas las variables en estudio un subconjunto de ellas que conforman el árbol de decisión C4.5 inducido, con la finalidad de generar el clasificador con esas variables seleccionadas y de esta forma, mejorar el poder predictivo (tasa de acierto) del clasificador y una simplificación del modelo.

Una vez obtenido el clasificador Naive Bayes preprocesado con el árbol de decisión C4.5 (NB-C4.5) se procedió a verificar su poder predictivo, para luego compararlo con los del clasificador Naive Bayes completo (NB-Completo). En la evaluación de los clasificadores se utilizaron tres conjuntos de datos provenientes del repositorio UCI, *Irvin Repository of Machine Learning databases de la Universidad de California* y un conjunto de datos proveniente de la Encuesta Nacional de Hogares del Instituto Nacional de Estadística e Informática del Perú, ENAHO – INEI, y es implementado con el programa WEKA [38].

4.1 Metodología

La metodología utilizada para verificar el poder predictivo de los clasificadores Naive Bayes C4.5 (NB-C4.5) y Naive Bayes completo (NB-Completo) en los diferentes conjuntos de datos, se detalla a continuación:

1. Seleccionar el subconjunto de variables que conformarán el clasificador Naive Bayes C4.5 (NB-C4.5) usando el algoritmo del árbol de decisión C4.5.
Para ello, procesar la base de datos mediante el algoritmo C4.5 y obtener el árbol de decisión que representa al conjunto de datos analizados. Las variables que componen dicha representación, pasaran a conformar el subconjunto de variables que serán tenidas en cuenta para obtener el clasificador Naive Bayes preprocesado (NB-C4.5).
2. Estimar el clasificador Naive Bayes con ese subconjunto de variables obtenido en (paso 1).
Esto es, estimar el clasificador Naive Bayes C4.5 (NB-C4.5) con las variables preseleccionadas.
3. Estimar el clasificador Naive Bayes completo (NB-Completo).
Esto es, estimar el clasificador Naive Bayes con todas las variables.
4. Comparar los clasificadores (NB-C4.5) y (NB-Completo) mediante su poder predictivo (tasa de acierto), utilizando la validación cruzada de 10 particiones.

En el presente estudio se está enfocando el clasificador Naive Bayes para variables discretas, pero en el caso de que se tenga variables continuas estas serán discretizadas antes de su uso. Se emplea la discretización usando intervalos con igual frecuencia con un valor de diez para el número de intervalos (ver Anexo).
La metodología será implementada con el programa WEKA.

En el Anexo, se muestra cómo obtener un árbol de decisión C4.5 y el clasificador Naive Bayes con el programa WEKA.

4.2 Evaluación del clasificador

La evaluación del clasificador se realizó en base a su poder predictivo (tasa de acierto), que es el número de casos clasificados correctamente dividido entre el número total de casos analizados.

$$tasa\ de\ acierto = \frac{casos\ clasificados\ correctamente}{número\ total\ de\ casos}$$

El poder predictivo del clasificador es basado en la tasa de acierto (o, a la inversa, la tasa de error) que posee.

Evaluar un clasificador, es importante no sólo para predecir su futuro comportamiento, sino también para poder escoger un clasificador (selección del modelo) dentro de un conjunto de posibilidades.

Para estimar el poder predictivo del clasificador se suele utilizar distintos métodos de validación. Los métodos de validación más usuales son, *hold-out* y *cross-validation* (validación cruzada).

El método *hold-out* es la técnica más sencilla, consiste en dividir la muestra disponible en dos conjuntos disjuntos de casos: uno para generar el modelo, *conjunto de entrenamiento* y otro para medir su bondad, *conjunto de prueba*, constituyendo generalmente 2/3 y 1/3 de la muestra, respectivamente. De esta forma, el conjunto de prueba es usado para estimar la tasa de acierto del clasificador.

Este método *hold-out* , se considera adecuado si el conjunto de datos es suficientemente grande, si no, la pérdida de información podría hacer que la estimación no fuera del todo fiable.

El método de validación cruzada (*cross-validation*), se basa en dividir en k conjuntos disjuntos del mismo tamaño. Cada uno de estos conjuntos será usado para probar el modelo obtenido con el resto de casos. El poder predictivo será estimado a partir de la media de las de tasas de acierto obtenidas en las k diferentes particiones.

Un caso particular del *cross-validation* es el dejar-uno-fuera (leave-one-out), se trata de separar un caso para la prueba y construir el clasificador con el resto de la muestra para clasificar éste. Esto se realizará con todos los casos de la muestra y el poder predictivo

será estimado como la proporción de casos correctamente clasificados con respecto al tamaño de la muestra.

En el presente trabajo, la estimación del poder predictivo de los clasificadores es obtenido mediante una validación cruzada de 10 particiones.

4.3 Aplicación a datos reales

Se comparó el poder predictivo de los clasificadores, Naive Bayes preprocesado con algoritmos de árbol de decisión C4.5 (NB-C4.5) y Naive Bayes completo (NB-Completo), utilizando tres conjuntos de datos provenientes del repositorio UCI , *Irvin Repository of Machine Learning databases de la Universidad de California* y un conjunto de datos proveniente de la Encuesta Nacional de Hogares del Instituto Nacional de Estadística e Informática del Perú, ENAHO – INEI, e implementado con el programa WEKA. Estos conjuntos de datos son presentados en el Cuadro 4.1, en ello se muestra la cantidad de casos, clases y variables.

Se adjunta al presente informe un disco compacto con los archivos de los datos.

Cuadro 4.1 .- Conjuntos de datos utilizados en la comparación de los clasificadores (NB-C4.5) y (NB-Completo)

Nº	Datos	Variables	Clases	Total casos
1	Iris	4	3	150
2	Vino	13	3	178
3	Cáncer (Winconsin)	9	2	683
4	Pobreza	9	3	6218

A continuación se presenta los diferentes conjuntos de datos, la descripción, variables y los resultados respectivos bajo la metodología propuesta.

4.3.1 CASO 1: Conjunto de datos IRIS

Este conjunto de datos IRIS es tal vez el más conocido que se encuentra en la literatura de clasificación. Usada por Fisher [10], y es un clásico en el campo y una referencia frecuente.

Descripción

El conjunto de datos de plantas Iris (IRIS), tiene un total de 150 casos. El objetivo es realizar una clasificación de plantas (lirios) a través de cuatro atributos de tipo continuo: el ancho y el largo tanto del pétalo como del sépalo. Hay un total de tres tipos de diferentes lirios (Iris Setosa, Iris Versicolor e Iris Virginica), ver Cuadro 4.2. Cada tipo de lirio presenta 50 casos.

Cuadro 4.2.- Variables del conjunto de datos IRIS

Variable	*Valor*	**Clase**
Longitud del sépalo (en cm)	Real	Tipo de lirio:
Ancho del sépalo (en cm)	Real	- *Setosa*
Longitud del pétalo (en cm)	Real	- *Versicolor*
Ancho del pétalo (en cm)	Real	- *Virgínica*

RESULTADOS

Para el conjunto de datos IRIS se obtuvo el árbol de decisión C4.5 mostrado en la Figura 4.1 . Se observa que las variables más representativas y que deben utilizarse en la obtención del clasificador Naive Bayes C4.5 (NB-C4.5) son dos: el ancho del pétalo y la longitud del pétalo.

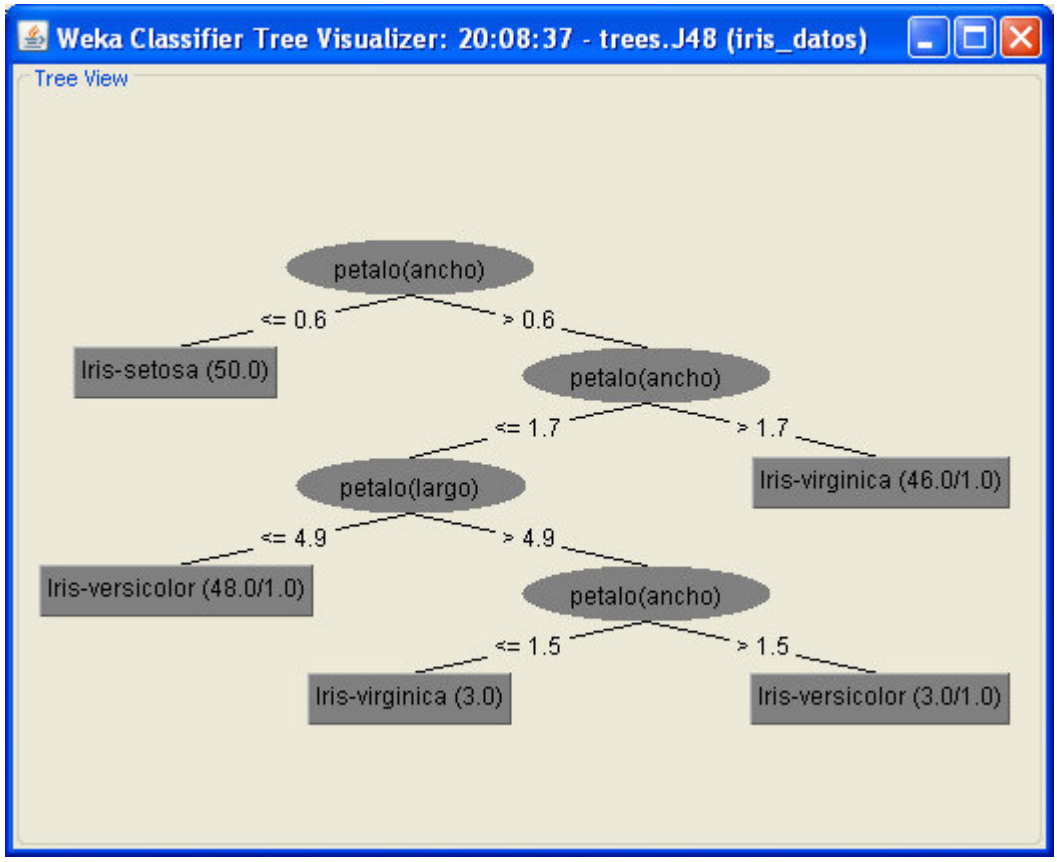


Figura 4.1.- Árbol de decisión C4.5 para la preselección de variables del conjunto de datos IRIS

Estas dos variables seleccionadas se utilizaron en la estimación del clasificador Naive Bayes C4.5 (NB-C4.5). En la Figura 4.2, se muestra la estructura del clasificador (NB-C4.5) para el tipo de planta (Clase) con las variables resultantes.

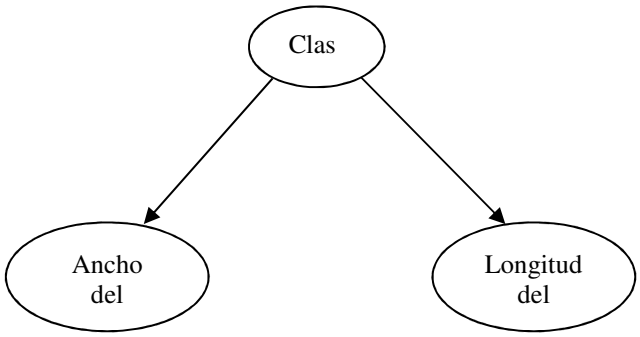


Figura 4.2.- Estructura del clasificador Naive Bayes C4.5 (NB-C4.5) a partir del conjunto de datos IRIS

Se puede decir, que estas variables seleccionadas son las más informativas para la clasificación del tipo de planta en uno de los tipos (Setosa, Versicolor o Virginica), esto es, las variables que mejor discriminan el tipo de planta.

En el Cuadro 4.3, se muestra el valor estimado del poder predictivo (porcentaje de aciertos) de los clasificadores NB-C4.5 y NB-Completo, y el número de variables utilizados en los respectivos clasificadores, en el análisis del conjunto de datos IRIS.

Cuadro 4.3.- Poder predictivo (porcentaje de aciertos) y número de variables de los clasificadores NB-C4.5 y NB-Completo del conjunto de datos IRIS

CLASIFICADOR	**Porcentaje de aciertos**	**Número de variables**
NB-C4.5	92.67%	2
NB-Completo	91.33%	4

Se observa que el clasificador propuesto NB-C4.5 obtuvo un porcentaje de aciertos (92.67%) superior al del clasificador NB-Completo (91.33%). Este clasificador NB-C4.5 además permitió la reducción del número de variables que inicialmente eran 4 en 2 variables.

El clasificador (NB-C4,5) ofrece un mejor rendimiento (porcentaje de aciertos) que el clasificador (NB-Completo), además el clasificador (NB-C4.5) no contiene a todas las variables predictoras, por tanto la complejidad del modelo puede verse reducida. Por ello se ve que, el clasificador (NB-C4.5) mejora el poder predictivo al compararse con el del clasificador (NB-Completo) y también puede contribuir a una simplificación del modelo. Además, la reducción en la complejidad puede facilitar la interpretación del modelo.

En el estudio de clasificación de plantas, podemos decir que, el ancho y la longitud del pétalo son variables que permiten diferenciarlos según su tipo de planta.

4.3.2 CASO 2: Conjunto de datos VINO

Este conjunto de datos de VINO es bastante utilizado en múltiples trabajos dentro del área de la clasificación. Muchos investigadores lo utilizan para comparar varios clasificadores.
En el contexto de clasificación, es un buen problema para probar un nuevo clasificador.

Descripción

Estos datos son el resultado de un análisis químico de los vinos producidos en una región de Italia pero que provienen de tres diferentes variedades. El análisis determinó las cantidades de 13 componentes que se encuentran en cada uno de los tres tipos de vinos (ver Cuadro 4.4).
Con respecto al número de casos por tipo de vino, se tiene: el vino tipo 1 con 59 casos representando el 33%, el tipo 2 con 71 casos (40%) y el tipo 3 con 48 casos representando un 27%. En total 178 casos.
Lo que se busca es clasificar el vino en uno de los tipos, teniendo la información de las variables especificadas.

Cuadro 4.4.- Variables del conjunto de datos VINO

Variable	*Valor*	**Clase**
Alcohol	Real	Tipo de vino
Acido málico	Real	- *Tipo 1*
Ceniza	Real	- *Tipo 2*
Alcalinidad de la ceniza	Real	- *Tipo 3*
Magnesio	Real	
Fenoles totales	Real	
Flavonoides	Real	
Fenoles no flavonoides	Real	
Proanthocyanins	Real	
Intensidad del color	Real	
Matiz	Real	
OD280/OD315 de vinos diluidos	Real	
Proline	Real	

RESULTADOS

Para el conjunto de datos de VINO se obtuvo el árbol de decisión C4.5 (Figura 4.3). Se observa que las variables más representativas y que deben utilizarse en la obtención del clasificador naive bayes C4.5 (NB-C4.5) son tres: Flavonoides, Prolina e Intensidad del color.

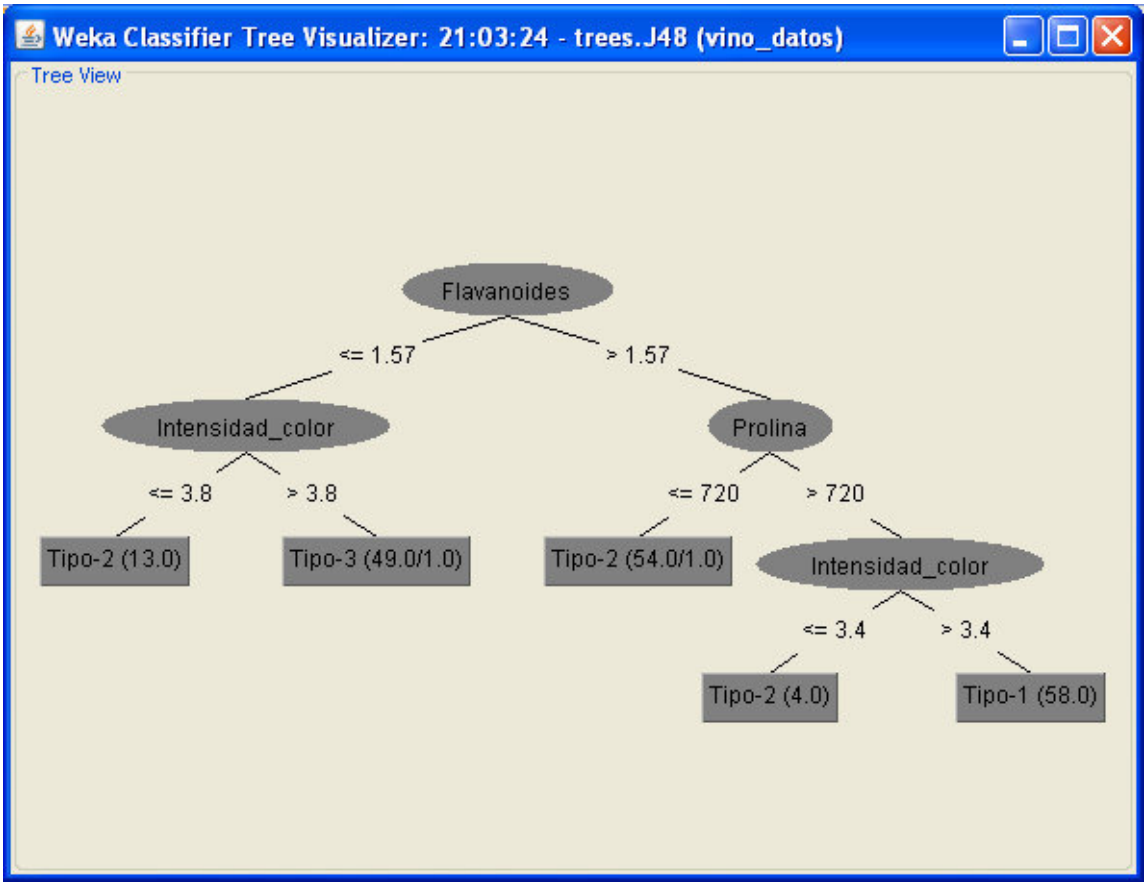


Figura 4.3.- Árbol de decisión C4.5 para la preselección de variables del conjunto de datos VINO

Estas tres variables seleccionadas se utilizaron en la estimación del clasificador Naive Bayes C4.5 (NB-C4.5). En la Figura 4.4 se muestra la estructura del clasificador (NB-C4.5) para el tipo de vino (Clase) con las variables resultantes.

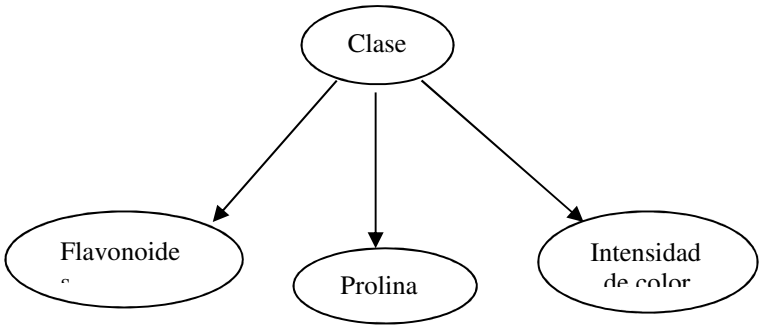


Figura 4.4.- Estructura del clasificador Naive Bayes C4.5 (NB-C4.5) a partir del conjunto de datos VINO

Se puede decir, que estas variables seleccionadas son las más informativas para la clasificación del vino en uno de los tipos (1, 2 o 3), esto es, las variables que mejor discriminan el tipo de vino.

En el Cuadro 4.5, se muestra el valor estimado del poder predictivo (porcentaje de aciertos) de los clasificadores NB-C4.5 y NB-Completo, y el número de variables utilizados en los respectivos clasificadores, en el análisis del conjunto de datos VINO.

Cuadro 4.5.- Poder predictivo (porcentaje de aciertos) y número de variables de los clasificadores NB-C4.5 y NB-Completo del conjunto de datos VINO

CLASIFICADOR	Porcentaje de aciertos	Número de variables
NB-C4.5	94.38%	3
NB-Completo	97.19%	13

Se puede apreciar que el clasificador propuesto NB-C4.5 obtuvo un porcentaje de aciertos (94.38%) inferior que el clasificador NB-Completo (97.19%). Este clasificador NB-C4.5 permitió la reducción del número de variables que inicialmente eran 13 en 3 variables.

A pesar de que el rendimiento (porcentaje de aciertos) del clasificador (NB-C4.5) no es mejor que el clasificador (NB-Completo), se debe tener en cuenta que en el clasificador NB-C4.5 no aparecen todas las variables predictoras y por tanto la complejidad del modelo puede verse reducida considerablemente. Por consiguiente, en un conjunto de datos con gran número de variables y para el cual la estimación de un modelo con todas ellas puede ser realmente costoso, el clasificador (NB-C4.5) puede contribuir a una simplificación del modelo sin tener porque sacrificar excesivamente la precisión en la clasificación. Además, esta simplificación del modelo también contribuye a una mayor facilidad para su interpretación.

En el estudio de clasificación de vinos, podemos decir que, la concentración de flavonoides, prolina e intensidad de color son variables que permiten diferenciarlos según su tipo de variedad (1, 2, o 3).

4.3.3 CASO 3: Conjunto de datos CANCER

Los datos se han obtenido de la base de datos de pacientes con cáncer de mama de la Universidad de Wisconsin. Se trata de una base de datos muy utilizada entre los científicos que estudian esta enfermedad y su diagnóstico.
El conjunto de datos CANCER de mama de Wisconsin es una de las más conocidas y utilizadas para probar algoritmos de clasificación de patrones de cáncer de mama.

Descripción

El conjunto de datos CANCER, contiene los datos de cáncer de mamá de 699 pacientes con 9 variables observadas vinculadas a esta patología y se clasifica el tipo de cáncer en dos tipos: benigno o maligno (ver Cuadro 4.6).

Se asignan valores enteros de 1 a 10 a las evaluaciones, siendo 1 el más cercano a benigno y 10 el más cercano a maligno.
Este conjunto de datos contiene 16 casos con valores de atributos perdidos que en este estudio se excluyeron del análisis. La base de datos contiene 444 (65,0%) casos con cáncer benigno y 239 (35,0%) malignos. En total se tienen 683 casos para el estudio.

Cuadro 4.6.- Variables del conjunto de datos CANCER

Variable	*Valor*	**Clase**
Espesor de la masa o grupo de células	1-10	Tipo de cáncer
Uniformidad del tamaño celular	1-10	- *Benigno*
Uniformidad de la forma celular	1-10	- *Maligno*
Adhesión marginal	1-10	
Tamaño individual de la célula	1-10	
Núcleo desnudo	1-10	
Cromatina blanda	1-10	
Núucleoli normal	1-10	
Mitosis	1-10	

RESULTADOS

Para el conjunto de datos CANCER se obtuvo el árbol de decisión C4.5 mostrado en la Figura 4.5 . Se observa que las variables más representativas y que deben utilizarse en la obtención del clasificador Naive Bayes C4.5 (NB-C4.5) son cinco: la uniformidad del tamaño celular, núcleo desnudo, uniformidad de la forma celular, espesor de la masa celular, adhesión marginal y cromatina blanda.

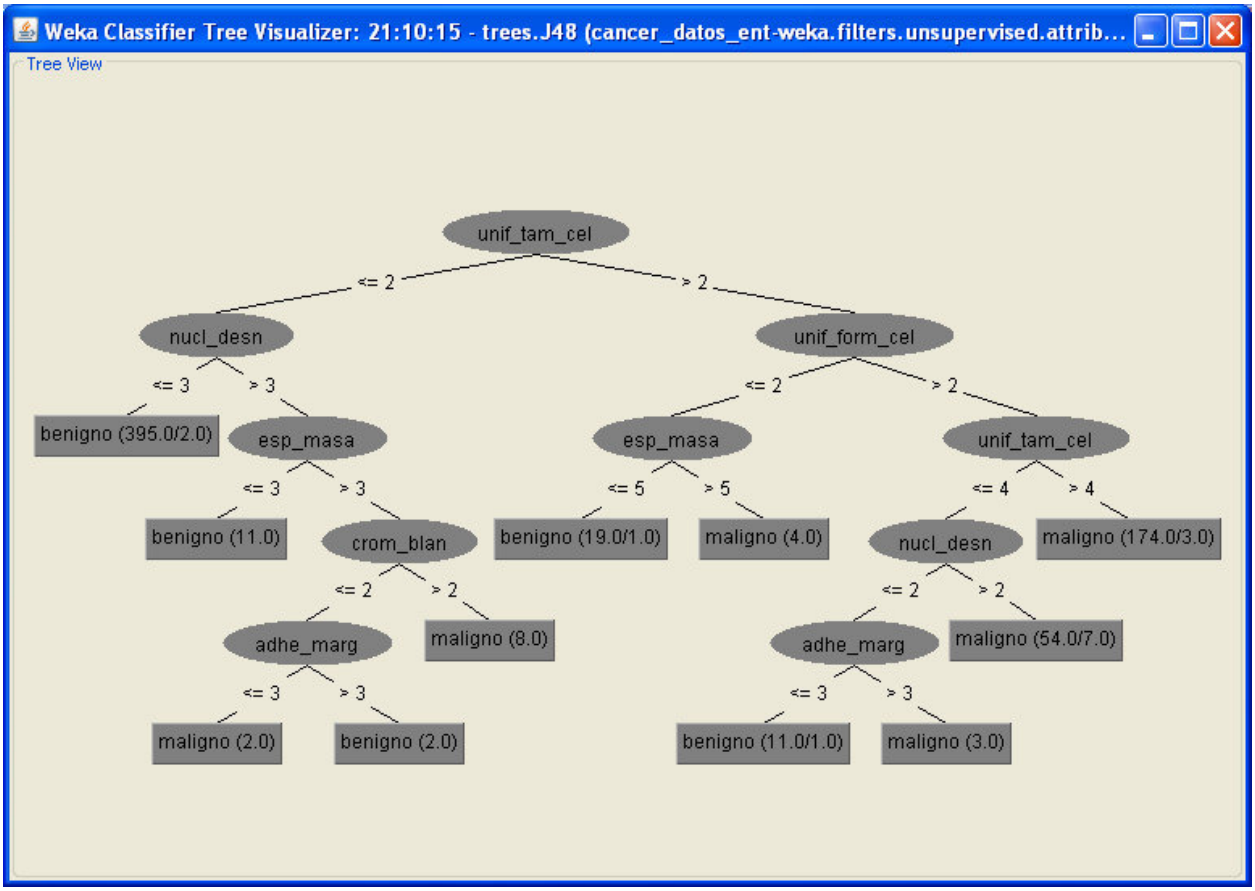


Figura 4.5.- Árbol de decisión C4.5 para la preselección de variables del conjunto de datos CANCER

Estas cinco variables seleccionadas se utilizaron en la estimación del clasificador Naive Bayes C4.5 (NB-C4.5). En la Figura 4.6, se muestra la estructura del clasificador (NB-C4.5) para el tipo de cáncer (Clase) con las variables resultantes.

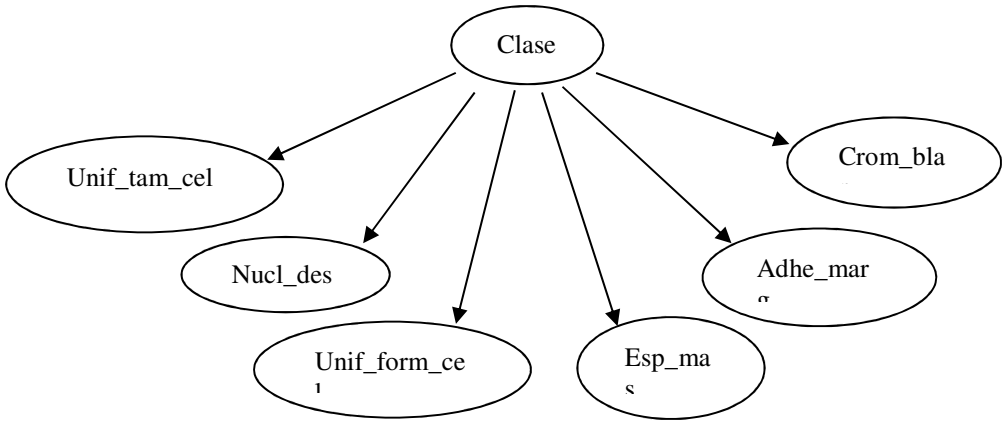


Figura 4.6.- Estructura del clasificador Naive Bayes C4.5 (NB-C4.5) a partir del conjunto de datos CANCER

Se puede decir, que estas variables seleccionadas son las más informativas para la clasificación del tipo de cáncer (Maligno o Benigno), esto es, las variables que mejor discriminan el tipo de cáncer.

En el Cuadro 4.7, se muestra el valor estimado del poder predictivo (porcentaje de aciertos) de los clasificadores NB-C4.5 y NB-Completo, y el número de variables utilizados en los respectivos clasificadores, en el análisis del conjunto de datos CANCER.

Cuadro 4.7.- Poder predictivo (porcentaje de aciertos) y número de variables de los clasificadores NB-C4.5 y NB-Completo del conjunto de datos CANCER

CLASIFICADOR	Porcentaje de aciertos	Número de variables
NB-C4.5	97.22%	6
NB-Completo	97.36%	9

Se observa que el clasificador propuesto NB-C4.5 obtuvo un porcentaje de aciertos (97.22%) aproximadamente igual al del clasificador NB-Completo (97.36%). Este clasificador NB-C4.5 permitió la reducción del número de variables que inicialmente eran 9 en 6 variables.

El clasificador (NB-C4,5) ofrece aproximadamente el mismo rendimiento (porcentaje de aciertos) que el clasificador (NB-Completo), además el clasificador (NB-C4.5) no contiene a todas las variables predictoras, por tanto la complejidad del modelo puede verse reducida de manera considerable. Por ello el clasificador (NB-C4.5) puede contribuir a una simplificación del modelo manteniendo el poder predictivo de clasificación. Además, la reducción en la complejidad puede facilitar la interpretación del modelo.

En el estudio de clasificación del tipo de cáncer, podemos decir que, los niveles de las variables: uniformidad del tamaño celular, núcleo desnudo, uniformidad de la forma celular, espesor de la masa celular, adhesión marginal y cromatina blanda permiten diferenciarlos según un cáncer benigno o maligno.

4.3.4 CASO 4: Conjunto de datos POBREZA

Este conjunto de datos POBREZA proviene de la Encuesta Nacional de Hogares (ENAHO - 2013) que es la investigación que permite al Instituto Nacional de Estadística e Informática (INEI) efectuar el seguimiento de los indicadores sobre las condiciones de vida y pobreza [18], [19].

Descripción

El conjunto de datos de hogares (POBREZA) contiene los datos de los indicadores sobre las condiciones de vida y pobreza de los hogares de la Región Selva del Perú. Se tiene un total de 6218 casos.

Para realizar la clasificación de los hogares respecto a su condición de pobreza se escogieron 9 atributos vinculados a la pobreza [18], [19]. La condición de pobreza es de tres tipos: Pobre extremo, Pobre no extremo y No pobre (ver Cuadro 4.8).

El conjunto de datos contiene 322 hogares con pobreza extrema, 1135 con pobreza no extrema y 4761 hogares no pobres.

Cuadro 4.8.- Variables del conjunto de datos POBREZA-ENAHO

Variable	*Valor*	**Clase**
Estrato geográfico (ESTRATO)	1 - 8	Condición de pobreza:
Total de perceptores de ingresos (PERCEPHO)	Entero	*- Pobre extremo*
Total de miembros del hogar (MIEPERHO)	Entero	*- Pobre no extremo*
Ingreso monetario (bruto) (INGMO1HD)	Real	*- No pobre*
Ingreso monetario (neto) (INGMO2HD)	Real	
Ingreso bruto (INGHOG1D)	Real	
Ingreso neto total (INGHOG2D)	Real	
Gasto monetario (GASHOG1D)	Real	
Gasto total bruto (GASHOG2D)	Real	

La variable Estrato Geográfico de los hogares tiene valores de 1 a 8, que indican si los hogares pertenecen a centros poblados con mucha o poca cantidad de viviendas, tal como:

Estrato Geográfico

1 Centros poblados mayor de 100,000 viviendas.
2 Centros poblados de 20,001 a 100,000 viviendas.
3 Centros poblados de 10,001 a 20,000 viviendas.
4 Centros poblados de 4,001 a 10,000 viviendas.
5 Centros poblados de 401 a 4,000 viviendas.
6 Centros poblados con menos de 401 viviendas.
7 Area de empadronamiento rural compuesta - AER Compuesto.
8 Area de empadronamiento rural simple - AER Simple.

RESULTADOS

Para el conjunto de datos POBREZA se obtuvo el árbol de decisión C4.5 mostrado en la Figura 4.7 . Se observa que el árbol es bastante profundo y frondoso, en este caso se utiliza las variables que se ubican en los primeros niveles, entonces, las variables más representativas y que deben utilizarse en la obtención del clasificador Naive Bayes C4.5 (NB-C4.5) son tres: GASHOG2D (Gasto total bruto), MIEPERHO (Total de miembros del hogar) y ESTRATO (Estrato geográfico).

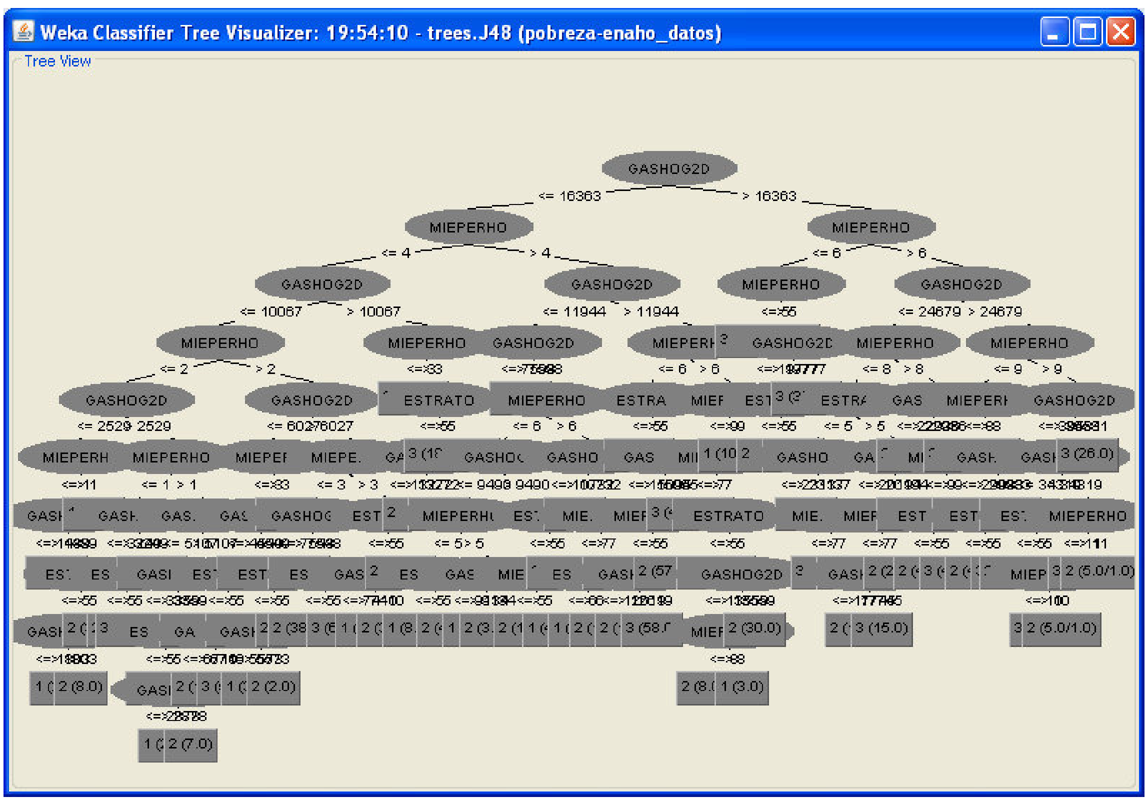


Figura 4.7.- Árbol de decisión C4.5 para la preselección de variables del conjunto de datos POBREZA

Estas tres variables seleccionadas se utilizaron en la estimación del clasificador Naive Bayes C4.5 (NB-C4.5). En la Figura 4.8, se muestra la estructura del clasificador (NB-C4.5) para la condición de pobreza (Clase) con las variables resultantes.

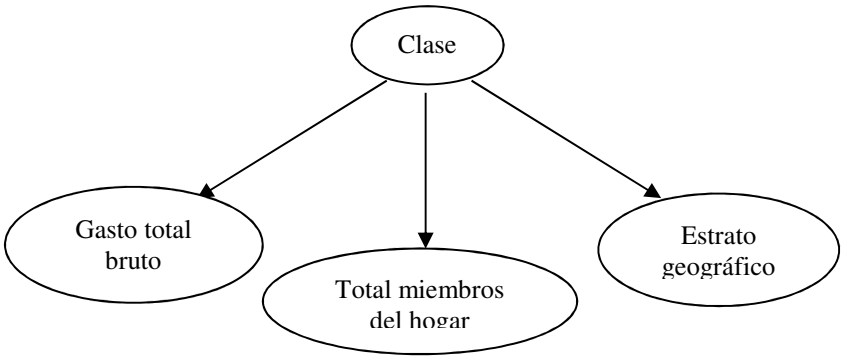


Figura 4.8.- Estructura del clasificador Naive Bayes C4.5 (NB-C4.5) a partir del conjunto de datos POBREZA

Se puede decir, que estas variables seleccionadas son las más informativas para la clasificación de los hogares en la Región Selva según su condición de pobreza (Pobre extremo, Pobre no extremo o No pobre), esto es, las variables que mejor discriminan la condición de pobreza de los hogares.

En el Cuadro 4.9, se muestra el valor estimado del poder predictivo (porcentaje de aciertos) de los clasificadores NB-C4.5 y NB-Completo, y el número de variables utilizados en los respectivos clasificadores, en el análisis del conjunto de datos POBREZA.

Cuadro 4.9.- Poder predictivo (porcentaje de aciertos) y número de variables de los clasificadores NB-C4.5 y NB-Completo del conjunto de datos POBREZA

CLASIFICADOR	Porcentaje de aciertos	Número de variables
NB-C4.5	83.37%	3
NB-Completo	66.85%	9

Se observa que el clasificador propuesto NB-C4.5 obtuvo un porcentaje de aciertos (83.37%) superior al del clasificador NB-Completo (66.85%). Este clasificador NB-C4.5 además permitió la reducción del número de variables que inicialmente eran 9 en 3 variables.

El clasificador (NB-C4,5) ofrece un mejor rendimiento (porcentaje de aciertos) que el clasificador (NB-Completo), además el clasificador (NB-C4.5) no contiene a todas las variables predictoras, por tanto la complejidad del modelo puede verse reducida. Por ello se ve que, el clasificador (NB-C4.5) mejora el poder predictivo al compararse con el del clasificador (NB-Completo) y también puede contribuir a una simplificación del modelo. Además, la reducción en la complejidad puede facilitar la interpretación del modelo.

En el estudio de clasificación de hogares, se puede decir que, el gasto total bruto anual, el total de miembros del hogar y el estrato geográfico son variables que permiten diferenciarlos según la condición de pobreza.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

En este presente trabajo se obtuvo un mejoramiento del clasificador Naive Bayes utilizando un método de preselección de las variables, que removió variables redundantes y/o irrelevantes del conjunto de datos con el uso del árbol de decisión C4.5 .

Al comparar el poder predictivo de los clasificadores, Naive Bayes propuesto (NB-C4.5) y Naive Bayes sin remover (NB Completo), se obtuvo que, para el conjunto de datos de la planta IRIS, el clasificador propuesto NB-C4.5 presentó un porcentaje de aciertos (92.67%) superior al del clasificador NB-Completo (91.33%), mientras que para el conjunto de datos de VINO, el clasificador propuesto NB-C4.5 obtuvo un porcentaje de aciertos (94.38%) inferior al del clasificador NB-Completo (97.19%), para el conjunto de datos CANCER, el clasificador propuesto NB-C4.5 obtuvo un porcentaje de aciertos (97.22%) aproximadamente igual al del clasificador NB-Completo (97.36%) y en el caso del conjunto de datos de POBREZA, el clasificador propuesto NB-C4.5 obtuvo un porcentaje de aciertos (83.37%) bastante superior del clasificador NB-Completo (66.85%).

Con respecto a la cuantificación de la reducción del número de variables obtenido por el clasificador Naive Bayes propuesto (NB-C4.5), se tiene que, para el conjunto de datos de la planta IRIS, éste clasificador permitió la reducción del número de variables que inicialmente eran 4 en 2 variables, mientras que para el conjunto de datos de VINO, éste clasificador permitió la reducción del número de variables que inicialmente eran 13 a 3 variables, para el conjunto de datos CANCER, el clasificador NB-C4.5 permitió la reducción del número de variables que inicialmente eran 9 en 6 variables y en el caso del

conjunto de datos POBREZA. el clasificador NB-C4.5 permitió la reducción del número de variables que inicialmente eran 9 en 3 variables.

Mediante la experimentación con los conjuntos de datos, se ha visto que, en la mayoría de los casos se tiene una mejora del poder predictivo del clasificador NB-C4.5 con respecto al del clasificador NB-Completo y además se obtiene una reducción del modelo que hace que sea más accesible a manejar. Por otra parte, la reducción en la complejidad también puede facilitar interpretabilidad de los modelos.

En resumen, basándonos en los resultados experimentales obtenidos concluimos que el método propuesto de preselección de variables mediante un árbol de decisión C4.5 mejora el rendimiento del clasificador Naive Bayes.

RECOMENDACIONES

Cuando se desea mejorar el rendimiento del clasificador Naive Bayes se recomienda utilizar el método propuesto, consistente en una preselección de variables, que busca remover variables redundantes y/o irrelevantes del conjunto de datos con el uso del árbol de decisión C4.5 para luego aplicar el clasificador Naive Bayes con esas variables obtenidas.

En tareas de clasificación se recomienda el clasificador Naive Bayes, ya que es un método muy efectivo, debido a su simplicidad y alto poder predictivo.

BIBLIOGRAFIA

[1] Ancell, R. (2013). Aportaciones de las redes bayesianas en meteorología. Predicción probabilística de precipitación. Tesis Doctoral. Universidad de Cantabria.

[2] Bala, J., Chang, K. C., Williams A. y Weng, Y. (2003). *A Hybrid Bayesian Decision Tree for Classification*. Workshop on Probabilistic Graphical Models for Classification, Cavtat-Dubrovnik, Croatia.

[3] Bouckaert, R. (1994). Probabilistic network construction using the minimum description length principle. Relatorio técnico. Utrecht University.

[4] Breiman, L., Friedman, J. H., Olshen, R. A. y Stone, C. J. (1984), Classification and Regression Trees, Wadsworth International Group.

[5] Chun-Nan Hsu, Hung-Ju Huang, y Tzu-Tsung Wong. (2000). Why discretization works for naive bayesian classifiers. In ICML ’00: Proceedings of the Seventeenth International Conference on Machine Learning, pages 399-406, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[6] Cooper, G., y Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9, 309-347.

[7] Díaz, L., Pacheco, S. y Garcia, R. (2004): El clasificador Naive Bayes para la extracción de conocimiento en bases de datos, Revista Ingenierías, No. 27.

[8] Domingos, P., y Pazzani, M. (1997). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning* 29:103-130.

[9] Dougherty, J., Kohavi, R. y Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the 12th International Conference on Machine Leaming,* pages 194-202.

[10] Fisher, R. A. (1936), 'The Use of Multiple Measurements in Taxonomic Problems', Annual Eugenics 7, 179–188.

[11] Felgaer, P. E. (2005). Tesis: Optimización de redes bayesianas basado en técnicas de aprendizaje por inducción. Facultad de Ingeniería, Universidad de Buenos Aires.

[12] Friedman, N. y Goldszmidt, M. (1996). Discretizing Continuous Attributes While Learning Bayesian Networks. In Internation Conference on Machine Learning.

[13] Friedman, N., Geiger, D. y Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29:2*, 131–163.

[14] Geiger, D. y Heckerman, D. (1994). Learning gaussian network. In Proccedings of the Tenth Conference on Uncertainty in Artificial Intelligence, pp. 235-243. Motrgan Kaufmann Publishers.

[15] Good. I. J. (1965). The estimation of probabilities. The MIT Press.

[16] Heckerman, D., Geiger, D. y Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20(3):197–243.

[17] Hunt, E. B., Marin, J., y Stone, P. T. (1966). Experiments in Induction. Academic Press, New York.

[18] Instituto Nacional de Estadística e Informática, INEI. (2013). Encuesta Nacional de Hogares 2013. Condiciones de vida y pobreza. Diccionario de datos. Lima: Talleres Gráficos INEI.

[19] Instituto Nacional de Estadística e Informática, INEI. (2013). Encuesta Nacional de Hogares 2013. Metodología para la obtención de variables calculadas (Sumaria). Lima: Talleres Gráficos INEI.

[20] Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29 (2), 119-127.

[21] Keogh, E., y Pazzani, M. (1999). Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics* (pp. 225–230).

[22] Kerber, R. (1992). Chimerge: discretization for numeric attributes. In Proceedings of National conference on artificial intelligence, pages 123-128, Menko Park. AAAI Press.

[23] Kononenko, I. (1990). Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. Current Trends in Knowledge Acquisition.

[24] Kononenko, I. (1991). Semi-naive Bayesian classifier. In Y. Kodratoff (Ed.), *Proc. Sixth European Working Session on Learning* (pp. 206–219). Berlin: Springer-Verlag.

[25] Lauritzen, S. y Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society series B, 50(2):157-224.

[26] Langley, P., Iba, W. y Thomas, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the Tenth National Conference of Artificial Intelligence*. AAAI Press. 223-228.

[27] Langley, P. y S. Sage. (1994). Induction of selective Bayesian classifiers. In R. L´opez de Mantar´as & D. Poole (Eds.), *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 399–406). San Francisco, CA: Morgan Kaufmann.

[28] Neapolitan (2004). *Learning Bayesian Networks*. Prentice Hall.

[29] Pazzani, M. J. (1995). An iterative improvement approach for the discretization of numeric attributes in Bayesian classi¯ers. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining.

[30] Pazzani, M. J. (1996). Constructive induction of Cartesian product attributes. In ISIS: Information, Statistics and Induction in Science, pp. 66-77, Melbourne, Aust. Word Scientific.

[31] Pearl, J. (1988). Probabilistic reasoning in Intelligence Systems: Networks of Plausible Inference, Morgan Kaufmann, Los Altos, CA.

[32] Quinlan, J. Ross (1986). Induction of decision trees. Machine Learning, 1(1):81-106.

[33] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.

[34] Ratanamahatana, C.A. y Gunopulos, D. (2002) 'Selective Bayesian classifier: feature selection for the Naive Bayesian classifier using decision trees', *Proceedings of*

the 3rd International Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields, September, Bologna, Italy, pp.613–623.

[35] Rish, I. (2001). An empirical study of the naive Bayes classifier. IBM Research Report, Yorktown Heights.

[36] Sahami, M. (1996). Learning limited dependence Bayesian classifiers. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 334–338). Menlo Park, CA: AAAI Press.

[37] Singh, M. y G. M. Provan. (1995). Efficient Learning of Selective Bayesian Network Classifier. *International Conference on Machine Learning*. Philadelphia, PA., Computer and Information Science Department, University of Pennsylvania.

[38] Witten, I. y Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Segunda Edición. Elsevier, 2005.

[39] Ying Yang y Geoffrey I, Web. (2002). A comparative study of discretization methods for naïve bayes classifiers. In Proceedings of PKAW 2002: The 2001 Pacific Rim Knowledge Acquisition Workshop, pages 159-173.

[40] Zheng, Z., y Webb, G. I. (2000). Lazy learning of Bayesian rules. Machine Learning, 41, 53-84.

ANEXO

ÁRBOL DE DECISIÓN C4.5 Y EL CLASIFICADOR NAIVE BAYES CON WEKA

WEKA (*Waikato Environment for Knowledge Analysis*), es un programa de minería de datos, un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos.

WEKA se distribuye como programa de libre distribución desarrollado en Java. Está constituido por una serie de paquetes con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación, y visualización.

Preparación de los datos

Los datos de entrada a la herramienta, sobre los que operarán las técnicas, deben estar codificados en un formato específico, denominado *Attribute-Relation File Format* (extensión "arff"). La herramienta permite cargar los datos en tres soportes: fichero de texto, acceso a una base de datos y acceso a través de internet sobre una dirección URL de un servidor web. En nuestro caso trabajaremos con ficheros de texto. Los datos deben estar dispuestos en el fichero de la forma siguiente: cada caso en una fila, y con los atributos separados por comas. El formato de un fichero arff sigue la estructura siguiente:

```
@relation NOMBRE_RELACION
@attribute r1 real
@attribute r2 real ...
...
@attribute i1 integer
@attribute i2 integer
…
@attribute s1 {v1, v2,…,vm}
@attribute s2 {v1, v2,…,vn}
…
@data
DATOS
```

por tanto, los atributos pueden ser principalmente de dos tipos: numéricos de tipo real o entero (indicado con las palabra *real* o *integer* tras el nombre del atributo), y simbólicos o categóricos, en cuyo caso se especifican los valores posibles que puede tomar entre llaves.

Para ejemplificar se trabajará con el conjunto de datos Diagnóstico de lentes (lente.arff), como se ve a continuación:

```
@relation lente

@attribute edad             {joven,pre_presbiopico,presbiopico}
@attribute padecimiento     {hipermetrope,miope}
@attribute astigmatismo     {si,no}
@attribute lagrimeo         {reducido,normal}
@attribute tipo_lente_clase {ninguno,suave,duro}

@DATA
joven,hipermetrope,si,reducido,ninguno
joven,hipermetrope,si,normal,duro
joven,hipermetrope,no,reducido,ninguno
joven,hipermetrope,no,normal,suave
joven,miope,si,reducido,ninguno
joven,miope,si,normal,duro
joven,miope,no,reducido,ninguno
joven,miope,no,normal,suave
pre_presbiopico,hipermetrope,si,reducido,ninguno
pre_presbiopico,hipermetrope,si,normal,ninguno
pre_presbiopico,hipermetrope,no,reducido,ninguno
pre_presbiopico,hipermetrope,no,normal,suave
pre_presbiopico,miope,si,reducido,ninguno
pre_presbiopico,miope,si,normal,duro
pre_presbiopico,miope,no,reducido,ninguno
pre_presbiopico,miope,no,normal,suave
presbiopico,hipermetrope,si,reducido,ninguno
presbiopico,hipermetrope,si,normal,ninguno
presbiopico,hipermetrope,no,reducido,ninguno
presbiopico,hipermetrope,no,normal,suave
presbiopico,miope,si,reducido,ninguno
presbiopico,miope,si,normal,duro
presbiopico,miope,no,reducido,ninguno
presbiopico,miope,no,normal,ninguno
```

Como convertir un archivo XLSX a ARFF para su lectura en WEKA

1) Convertir el archivo **xlsx** a **csv** (Abrir **Excel** y dar guardar como .csv (delimitado por comas))
2) Ahora abrir el archivo **.csv** con el **Bloc de Notas**
3) En el **Bloc de Notas** reemplazar **“,”** por **“.”** (si es que el Excel ha detectado los datos: ejemplo el 1.15 con comas 1,5).
   ¿Cómo se hace? En la barra de herramientas pulsar **Edición / Reemplazar**, completamos (**Buscar: “,” / Reemplazar por: “.”**) y dar **Reemplazar Todo**.
4) En el **Bloc de Notas** reemplazar **“;”** por **“,”**
5) En el **Bloc de Notas** guardar como dándole un nombre con extensión **.arff** (por ejemplo *lente.arff*)
6) Abrir **Weka** y dar open file, seleccionar *lente.arff* , y desde ahí se puede trabajar

En el cuadro siguiente se muestra una parte del archivo de datos IRIS (**iris.arff**) que contiene datos con decimales:

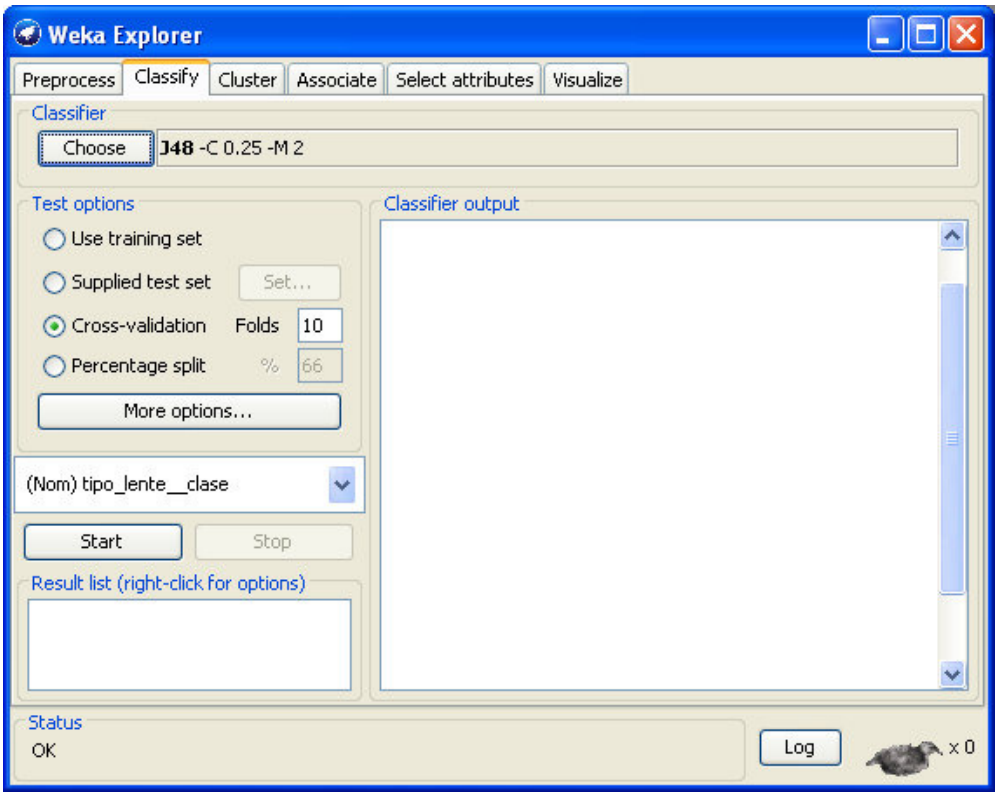
```
@relation iris_datos

@attribute sepalo(largo)       real
@attribute sepalo(ancho)       real
@attribute petalo(largo)       real
@attribute petalo(ancho)       real
@attribute clase {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
……
……
```

Generación del árbol de decisión C4.5 (J48 en Weka)

Se generará un árbol de decisión C4.5 (implementado en Weka como J48) con el conjunto de datos Diagnóstico de lentes.

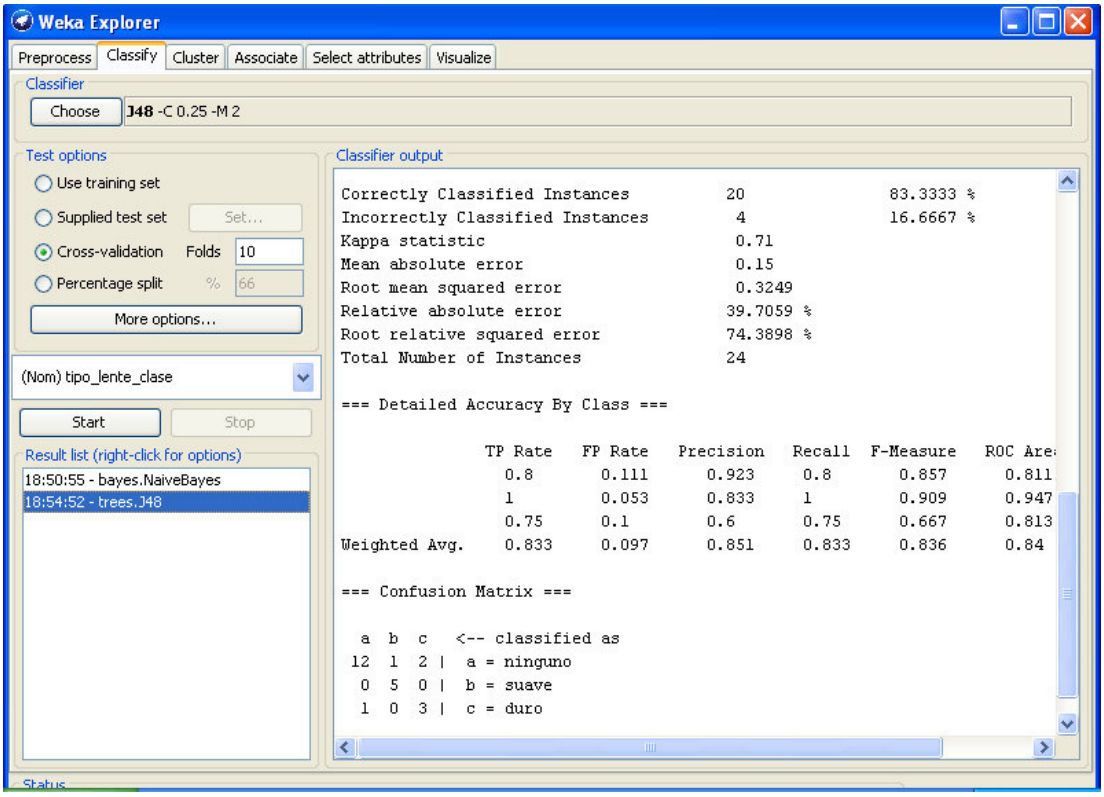
Abrir el programa Weka, elegir **Explorer** luego en la pestaña *Preprocess* pulsar **Open File…** y seleccionará el archivo a trabajar, en este caso, el conjunto de datos *lente.arff*.

Luego se selecciona la pestaña **Classify** y se elegirá un clasificador pulsando el botón **Choose**. Aparecerá una estructura de directorios en la que se seleccionará el directorio **trees** y dentro del él el algoritmo **J48**. Se mantendrán las opciones por defecto del clasificador (**J48 –C 0.25 –M 2**), donde 0.25 es el factor de confianza usado para la poda (con un menor valor se incurre en más poda) y 2 indica el mínimo número de casos por hoja), tal y como muestra la pantalla siguiente.



El resto de opciones para el experimento también se mantendrán en los valores por defecto: activa la opción de test '**cross validation**' e inactivas las restantes. Para generar el árbol se pulsará **Start**.

El resultado será el que muestra la pantalla siguiente, donde se muestran en modo texto tanto el árbol generado como la capacidad de clasificación del mismo:



Si se analiza la información que se ofrece en modo texto, se puede destacar lo siguiente:

En primer lugar, se muestra información sobre el tipo de clasificador utilizado (algoritmo **J48**), la base de datos sobre la que se trabaja (lente) y el tipo de test (**cross validation**).

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     lente
Instances:    24
Attributes:   5
              edad
              padecimiento
              astigmatismo
              lagrimeo
              tipo_lente_clase
Test mode:10-fold cross-validation
```

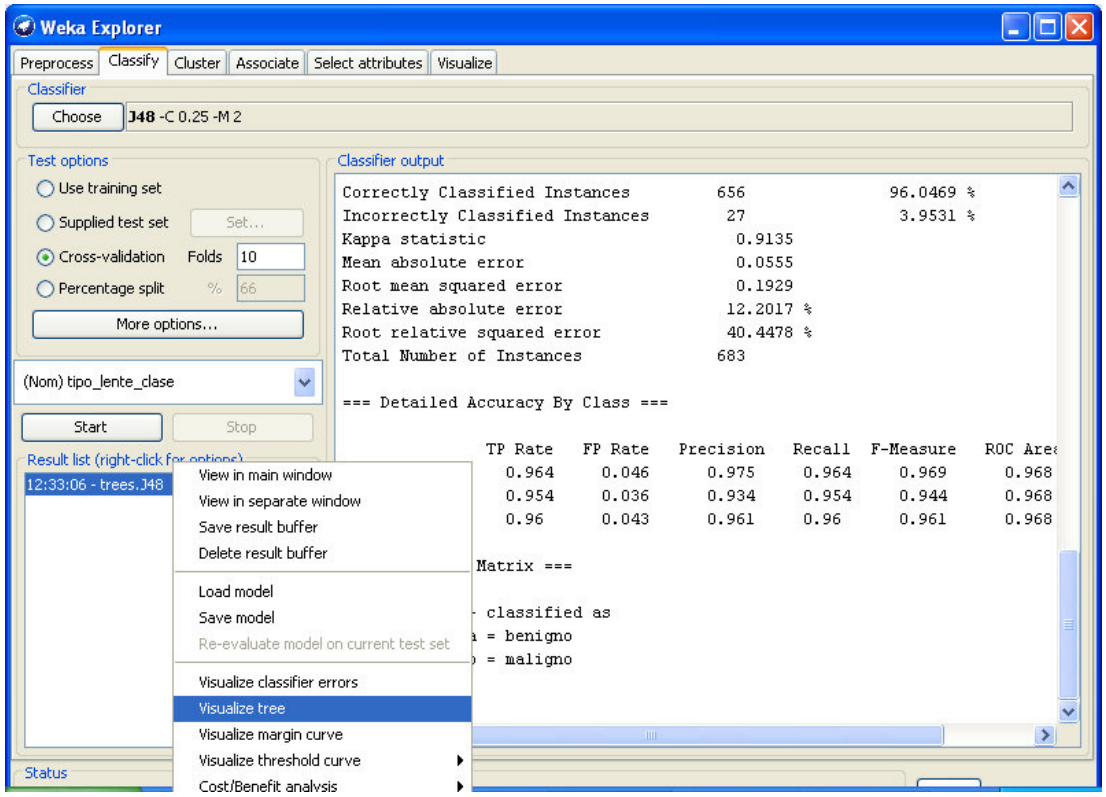
A continuación se muestra el árbol que se ha generado y el número de instancias que clasifica cada nodo:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

lagrimeo = reducido: ninguno (12.0)
lagrimeo = normal
|   astigmatismo = si
|   |   padecimiento = hipermetrope: ninguno (3.0/1.0)
|   |   padecimiento = miope: duro (3.0)
|   astigmatismo = no: suave (6.0/1.0)

Number of Leaves  :    4

Size of the tree :     7

Time taken to build model: 0.02 seconds
```

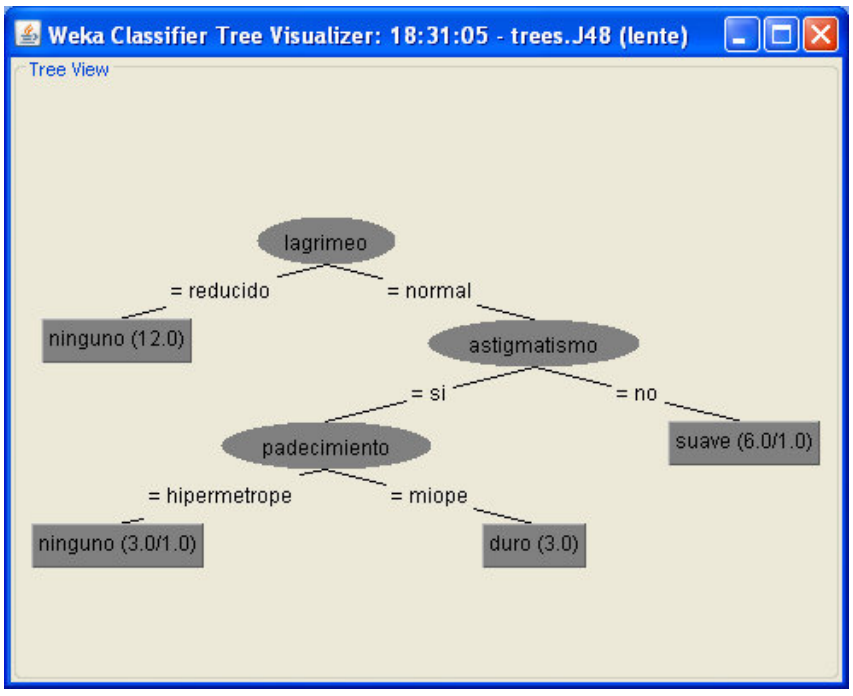
Y por último se muestran los resultados del test (indican la capacidad de clasificación esperable para el árbol y la matriz de confusión):

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          20               83.3333 %
Incorrectly Classified Instances         4               16.6667 %
Kappa statistic                          0.71
Mean absolute error                      0.15
Root mean squared error                  0.3249
Relative absolute error                 39.7059 %
Root relative squared error             74.3898 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  ROC Area  Class
               0.8       0.111     0.923       0.8     0.857      0.811     ninguno
               1         0.053     0.833       1       0.909      0.947     suave
               0.75      0.1       0.6         0.75    0.667      0.813     duro
Weighted Avg.  0.833     0.097     0.851       0.833   0.836      0.84

=== Confusion Matrix ===

  a  b  c   <-- classified as
 12  1  2 |  a = ninguno
  0  5  0 |  b = suave
  1  0  3 |  c = duro
```

También es posible visualizar el árbol de decisión de una forma más legible. Para ello se debe hacer clic con el botón derecho en la ventana de resultados, sobre el resultado de la generación del árbol. Aparecerá un menú desplegable:



Y dentro de ese menú se deberá seleccionar la opción **'Visualize tree'**. El resultado se muestra en la figura siguiente:



Generación del clasificador Naive Bayes con Weka

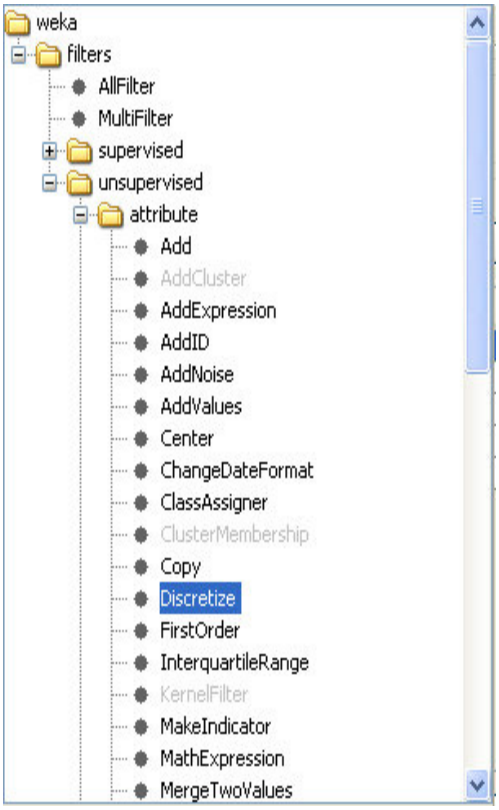
Se generará el clasificador Naive Bayes a partir del conjunto de datos Diagnóstico de lentes.

Abrir el programa Weka, elegir **Explorer** luego en la pestaña *Preprocess* pulsar **Open File…** y seleccionará el archivo a trabajar, en este caso, el conjunto de datos *lente.arff*.

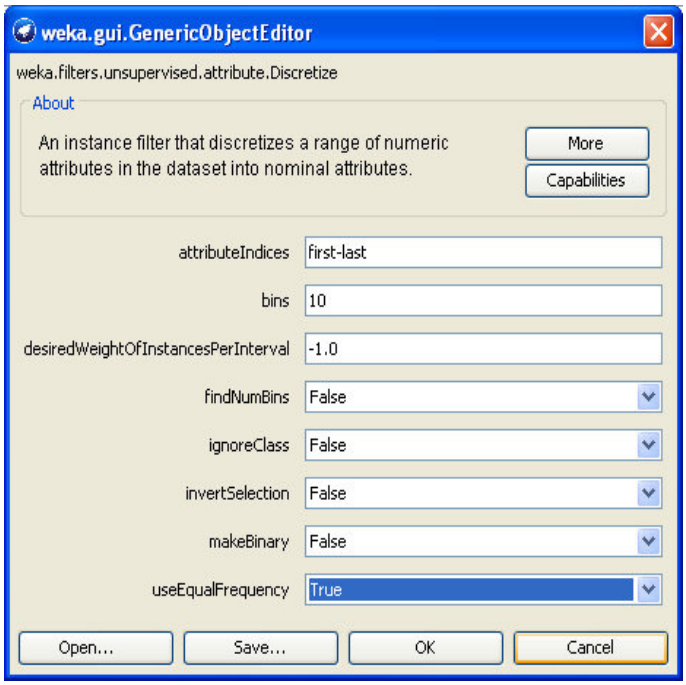
Notar que, si el conjunto de datos presentan variables continuas éstas tendrán que ser discretizadas (en este presente estudio se emplea la discretización usando intervalos con igual frecuencia con un valor de diez para el número de intervalos).

Weka presenta la herramienta que pertenece a la categoría de discretización no supervidasa y puede ser configurada para crear intervalos de igual frecuencia, la cual se encuentra siguiendo la ruta mostrada en las siguientes salidas:

Para ello, en la pestaña *Preprocess* elegir el botón **Choose** del apartado "Filter" siguiendo la ruta mostrada, llegando a la herramienta **Discretize**:



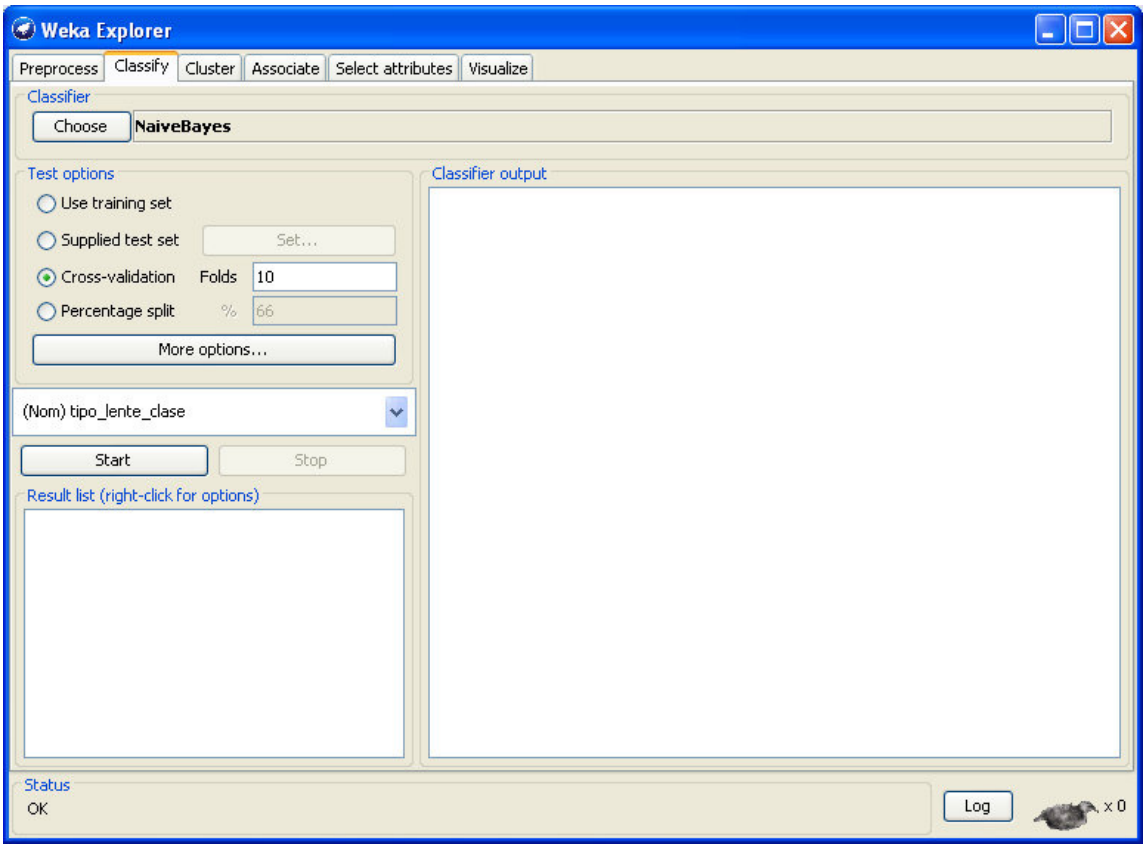
Esta herramienta implementa el algoritmo de discretización y es posible configurar sus propiedades haciendo clic con el botón derecho del ratón sobre ella una vez seleccionado, lo que despliega la ventana como se puede observar:



Para emplear la discretización usando intervalos con igual frecuencia con un valor de diez para el número de intervalos, es necesario asignar el valor "**True**" a la propiedad **useEqualFrequency** de la ventana de propiedades anterior, se puede observar que por defecto se presenta la opción de **bins 10** (10 intervalos).

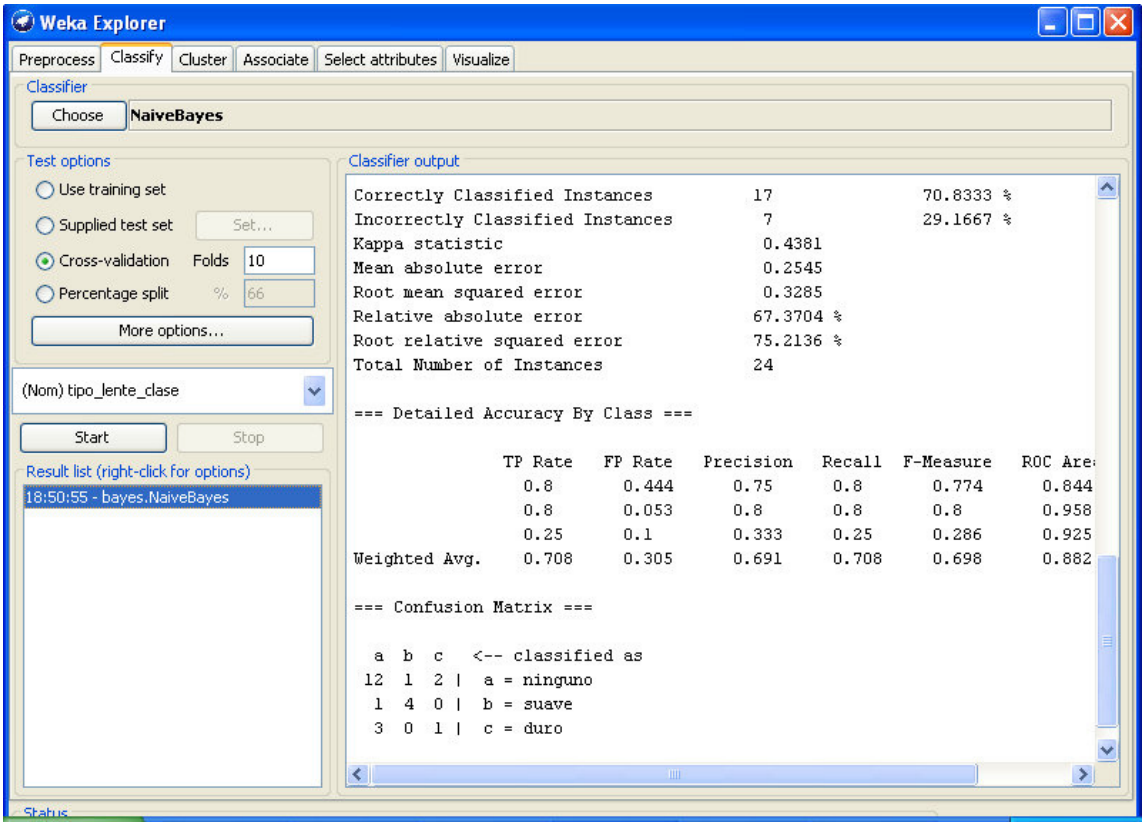
Después de la etapa de discretización (en este ejemplo no es necesario la discretización, ya que las variables no son continuas), se continua con la estimación del clasificador.

Seleccionar la pestaña **Classify** y elegir un clasificador pulsando el botón **Choose**. Aparecerá una estructura de directorios en la que se seleccionará el directorio **bayes** y dentro del él el algoritmo **NaiveBayes**, tal y como muestra la pantalla siguiente:



El resto de opciones para el experimento también se mantendrán en los valores por defecto: activa la opción de test '**cross validation**' e inactivas las restantes. Para generar el clasificador se pulsará **Start**.

El resultado será el que muestra la pantalla siguiente, donde se muestran en modo texto tanto las probabilidades del clasificador como la capacidad de clasificación del mismo:



Si se analiza la información que se ofrece en modo texto, se puede destacar lo siguiente:

En primer lugar, se muestra información sobre el tipo de clasificador utilizado (**NaiveBayes**), la base de datos sobre la que se trabaja (**lente**) y el tipo de test (**cross validation**).

```
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation:     lente
Instances:    24
Attributes:   5
              edad
              padecimiento
              astigmatismo
              lagrimeo
              tipo_lente_clase
Test mode:10-fold cross-validation
```

A continuación se muestra las probabilidades a priori de las clases y las frecuencias respectivas corregidas que se utilizan en el cálculo de las probabilidades condicionales (con la corrección de Laplace) para cada atributo dada la clase (tipo de lente):

=== Classifier model (full training set) ===

Naive Bayes Classifier

Attribute	Class ninguno (0.59)	suave (0.22)	duro (0.19)
edad			
joven	5.0	3.0	3.0
pre_presbiopico	6.0	3.0	2.0
presbiopico	7.0	2.0	2.0
[total]	18.0	8.0	7.0
padecimiento			
hipermetrope	9.0	4.0	2.0
miope	8.0	3.0	4.0
[total]	17.0	7.0	6.0
astigmatismo			
si	9.0	1.0	5.0
no	8.0	6.0	1.0
[total]	17.0	7.0	6.0
lagrimeo			
reducido	13.0	1.0	1.0
normal	4.0	6.0	5.0
[total]	17.0	7.0	6.0

Time taken to build model: 0 seconds

Y por último se muestran los resultados del test (indican la capacidad de clasificación esperable para el clasificador y la matriz de confusión):

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          17               70.8333 %
Incorrectly Classified Instances         7               29.1667 %
Kappa statistic                          0.4381
Mean absolute error                      0.2545
Root mean squared error                  0.3285
Relative absolute error                 67.3704 %
Root relative squared error             75.2136 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0.444      0.75      0.8       0.774      0.844    ninguno
                 0.8       0.053      0.8       0.8       0.8        0.958    suave
                 0.25      0.1        0.333     0.25      0.286      0.925    duro
Weighted Avg.    0.708     0.305      0.691     0.708     0.698      0.882

=== Confusion Matrix ===

  a  b  c   <-- classified as
 12  1  2 |  a = ninguno
  1  4  0 |  b = suave
  3  0  1 |  c = duro
```